The diffusion of innovations in the presence of geography and media.

by

Jameson Lawrence Toole

B.S. - Honors Physics, University of Michigan (2010)
B.S. - Honors Economics, University of Michigan (2010)
B.S., University of Michigan (2010)

Submitted to the Engineering Systems Division
in partial fulfillment of the requirements for the degree of

Masters of Science - Engineering Systems

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2012

Author . . . .
Engineering Systems Division
January 20, 2012

Certified by . . . .
Marta C. González
Assistant Professor
Thesis Supervisor

Accepted by . . . .
Olivier L. de Weck
Chairman, ESD Education Committee

The diffusion of innovations in the presence of geography and media.

by

Jameson Lawrence Toole

Submitted to the Engineering Systems Division
on January 20, 2012, in partial fulfillment of the
requirements for the degree of
Masters of Science - Engineering Systems

Abstract

Increasingly, the world we live in is digital, mobile, and online. As a consequence, many of your seemingly mundane actions are recorded, archived, and for the first time widely accessible to both the generators and curators of this information. From this fire hose of digital breadcrumbs, we can learn an enormous amount about ourselves as individuals and societies. Simple questions such as where we go, who we are meeting, and how we interact when we get there can be explored with incredibly high resolution and richness. Through new emiprical and analytic tools, we can leverage information generated from rapidly expanding online social networks, revealing the beautiful and often surprising complexity of everyday human behavior. We are able to harness data from millions of cell phone users to better understand how people move through cities, use roads, and interact with their neighbors.

This thesis deals with quantifying, analyzing, and ultimately modeling socio-technical systems. More specifically, it focuses on modeling the diffusion of innovations in time and space. While there has been much work examining the affects of social network structure on innovation adoption, models to date have lacked important features such as meta-populations reflecting real geography or influence from mass media forces. This thesis shows that these are features crucial to producing more accurate predictions of a social contagion and technology adoption at the city level. Using data from the adoption of the popular micro-blogging platform, Twitter, a model of adoption on a network is presented. The model places friendships in real geographic space and exposes individuals to mass media influence. Results show that homophily both amongst individuals with similar propensities to adopt a technology and geographic location is critical to reproduce features of real spatiotemporal adoption. Furthermore, estimates suggest that mass media was responsible for increasing Twitter's user base two to four fold. To reflect this strength, traditional contagion models are extended to include an endogenous mass media agent that responds to those adopting an innovation as well as influencing agents to adopt themselves.

The final chapter of this thesis addresses the future. The ubiquity of digital devices like mobile phones and tablets is opening rich new avenues of research. The massive amounts of data generated and stored by these devices can be used to gain a better understanding of the complex socio-technical systems they sense. The same tools, techniques, and analogies utilized in the first three chapters of this thesis can now literally be taken to the streets. With mobile phones that record when and where activities take place, a new window has been opened on urban systems. Future work will explore how people use cities dynamically to improve transportations systems and inform urban planners. New measurements will help understand what cities do well, when they fail, and why.

At the core of this new domain, is an interdisciplinary approach to complex socio-technical systems that combines many fields and methods. This view forms a more holistic view of problems and potential solutions. The thesis presented stands as an example of data, theory, and simulation for diverse areas can be combined to gain novel insights into human behavior.

Thesis Supervisor: Marta C. González
Title: Assistant Professor

Acknowledgments

First and foremost I would like to thank my advisor, Marta C. González, for her support, guidance, and patience throughout the duration of this research and thesis. Her generous offering of knowledge and advice has enabled me to navigate the challenging but rewarding process of completing this thesis. Under her tutelage, I have been able grow as a student and researcher while maintaining a genuine fascination and passion for my work and for this I owe her thanks and gratitude.

To my parents, I cannot thank you enough for your constant and unwavering encouragement, your willingness to listen, and your unconditional love. You have shown me what it means to work hard, seek happiness, and most of all, help others along the way. If nothing else, know that I was watching and that your strength and kindness has not gone unnoticed. I can only hope to pay forward the wonderful things you have given me.

I feel incredibly lucky and proud to have family and friends who do so much to inspire and include me. The breadth and depth of your accomplishments has served as a constant reminder to dream big, try my best, and have fun. I am truly honored to know you and cannot wait see what amazing things lie ahead. Finally, I cannot forget all of the nurturing and generous educators who have given me the knowledge and wonder that has carried me this far. I hope it is enough to know that you have made a difference for at least one student and for that I sincerely thank you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction and Overview

Complex socio-technical systems present numerous challenges to empirical study and modeling. Large and diverse sets of variables interact in ways that often lead to emergent phenomena at different scales. These problems are especially prevalent in systems involving human behavior and choice. Some ground is gained by assuming individuals are rational and forward thinking, but when these assumptions are relaxed, complexity quickly outpaces traditional modeling tools.

In the past decade, the proliferation of digital, mobile, and online technology has emerged, promising to foster new insights. These digital services and devices automatically record and store clicks, searches, calls, posts, tweets, and countless other daily events. The mountain of digital breadcrumbs generated by billions of individuals has spurred the growth of *computational social science* [29], a new approach dedicated to empirically measuring social phenomena on large of scales and with high resolution. This amazing sea of data promises to help us extract and understand fundamental laws of socio-technical systems. Moreover, the insight provided can be used to improve these systems, making them cheaper, more efficient, and sustainable.

This thesis focuses on the diffusion of innovation and information through spatial social networks. The diffusion of innovation is a universal and important process. Understanding how an idea spreads or a technology is adopted has been a fundamental question asked by industries, governments, and academics alike. Products, services, and ideas are useless unless people know about them. The difficulty, though, is in the process. Systems that facilitate the spreading of an innovation and the incentives for adopting it are complex. Individuals gain benefits from more efficient technologies, incur transmission costs for seeking them out, and are often influenced by a number of externalities from people and institutions around them. The relative weight of these forces and how they interact is still unknown.

Consider the fax machine. The first models were very expensive, making the cost of adoption high. Moreover, fax machines exhibit several externalities. The owner of the world's only fax machine is quite lonely because there is no one receive the fax. However, if every home and office is equipped with one, the machine becomes an invaluable communication tool. Similarly, there may be different standards and formats for faxing information. An early adopter risks choosing a format that becomes obsolete or incompatible. On top of these concerns, there may be mass media campaigns or government regulations that further complicate choices. The question, then, is how all of these forces combine to affect the resulting adoption process of the fax machine.

The implications of understanding these processes are broad. Companies may benefit from better prediction of the demand for their goods and services. This, in turn, may lead to more efficient production or lower costs. Governments and public institutions might increase the effectiveness and efficiency of policies and programs. Even individuals can flourish from faster exposure to more relevant information, ideas, and communities. All of these improvements, rest on better understanding how and with whom we communicate and how that communication influences actions and behaviors.

In an increasingly digital and connected world, the processes by which information is shared and consumed are changing rapidly. Services and content are now distributed through on-line social networks where the flattening affects of the internet distort diffusion in both space and time. Today, an email can inform a neighbor next door as fast as a friend on the opposite side of the globe. In addition to quicker communication, the institutions that communicate are also changing. Previously, there were few alternatives to spreading ideas through word-of-mouth. There are now strong mass media outlets, capable of reaching millions with a single broadcast. Very recently, communication channels have again been disrupted by the rise of online communities and applications. These factors are quickly shifting the balance between word-of-mouth contagion and more traditional mass media advertisement and are changing the spatiotemporal scales on which spreading occurs. Because of these shifting dynamics, new approaches are required to understand and ultimately forecast the diffusion of innovation through populations.

In addition to changing how we spread innovations, the incentives to share and use them are also transforming. Social applications and services have enormous externalities. They become more valuable as more users sign up to use them. Moreover, products and services in the information age are less limited by expensive production and transportation costs. The incentives to buy expensive, durable goods, like a new car, are very different from the those considered when deciding to sign up for a new web application. In the past, the focus was on the diffusion and adoption of high risk investments, while current investments come with lower costs and fewer risks. To incorporate these properties, models of innovation diffusion need to be updated.

Aiding our ability to characterize and quantify this shift are unprecedented amounts of data illuminating how people communicate with each other and how that communication translates into choices or behaviors like adopting an innovation. The very devices responsible for enabling this paradigm shift can be used as sensors to study and understand it. Current estimates show that nearly 30% of the world's population are internet users and an astounding 80% are mobile phone subscribers[1]. Using these connections and devices as sensors presents an opportunity to study the behaviors of huge sections of the population.

These data provide a window into the process of how information and innovation spread. Not only is it easier to track the level of adoption, but it is increasingly

[1] Statistical estimates made by the International Telecommunications Union, a UN agency. http://www.itu.int/dms_pub/itu-d/opb/ind/D-IND-ICTOI-2011-SUM-PDF-E.pdf

possible to measure who is adopting and why. Moreover, it is not just individuals using these new technologies. Traditional and new social media outlets are continually finding novel ways to converse and connect with content consumers. Communication patterns and content are broadcast on a the social web and these data are now available to download, store, and analyze. This thesis leverages the new paradigm in data to inform models of how innovation spreads and measure the relative importance of features such as word-of-mouth and mass media.

1.2 Thesis Structure

This thesis presents a study quantifying, analyzing, and ultimately modeling a socio-technical system. It explores the diffusion of innovation in a population across both space and time. The first chapter places this work in the context of important problems facing industries, government, and academics. The second chapter discusses previous work in the area. It covers the efforts of traditional social science and business to understand the phenomena of innovation diffusion as well as more recent work by the statistical physics community to model social dynamics.

The third chapter presents empirical analysis of real-world innovation diffusion followed by a model and simulation capturing salient features of reality. The model is grounded in the analysis of data about millions of users on the popular micro-blogging platform, Twitter. It quantifies the spread of Twitter throughout the United States, using adoption data from the first three years of Twitter's existence. The insights gained from analysis of this data are then used to build a model that extends and unifies previous frameworks in a way that more accurately reproduces features of real technology adoption. The results of the modeling simulations are compared and discussed. The broad conclusion of this work suggests that the geographic distribution of friendships is extremely important to replicate spatial diffusion of innovations and that mass media influence is responsible for more user adoption than word-of-mouth spreading.

The last chapter outlines some future directions for research. It explores the possibilities from layering traditional data about urban environments on top of novel data from digital devices such as mobile phones. The common thread between these two topics is the ability of extract and understand patterns of human behavior from digital data. Though potential phenomena of study occur on many different spatial and temporal scales, they demonstrate the breadth and depth of insights that these massive data sets offer. Furthermore, this work highlights the need to develop interdisciplinary approaches understanding complex socio-technical systems.

Chapter 2

Literature Review: Diffusion of Innovations and the Physics of Social Phenomena

2.1 Introduction

This chapter presents a range of literature addressing the diffusion of innovations and dynamics of social phenomena. A wide variety of researchers, from economists to physicists, have studied how and why groups of individuals choose to adopt a technology, share information, or participate in collective action. An equally wide variety of models have been proposed to explain the process. Some simulate the spread of a disease, others predict the purchase of a new type of crop. The main goal of this chapter is present past and current ways of modeling innovation diffusion, translate between parallel fields, and point out gaps to be addressed by the new model presented in Chapter 3.

2.2 Compartment Models

The oldest models of contagion focus on the spread of disease [14] or the diffusion of innovation [40, 46]. Simple approaches known as *compartment models* have proven extremely informative to epidemic modeling. The most notable of these compartment models is the susceptible - infected (SI) model. The diffusion of innovations literature has had made use of similar frameworks, such as the Bass model [5]. Compartment models allow individuals to occupy a single state (compartment), e.g. "infected by a disease". Transition dynamics are then specified to control how individuals may move between states. A brief overview of simple compartment models is presented before discussing more complex, agent based approaches in later sections.

2.2.1 Susceptible - Infected Models

The simplest contagion model is the susceptible - infected or SI model. The SI framework assumes a fixed population of N individuals, wherein each individual may occupy one compartment or state. An individual can either be *susceptible (S)* to the contagion or *infected (I)* by it. By conservation, the number of susceptible individuals, S, and the number of infected individuals, I, must add to the total population so that $S + I = N$. It is also common to generalize to the continuous case where all values are normalized by the population size to represent the fraction of individuals in each state. Finally, a population is initialized with nearly all agents in the susceptible state. However, a few infected individuals, known as the *seed*, must be present because there is no mechanism to spontaneously start the contagion process.

The dynamics of the SI framework specify how individuals transition from one state to another. In order to transmit a disease, two things must happen. First, two individuals (one infected, one susceptible) must come into contact with each other. Then, transmission of the disease must occur from the infected to the susceptible individual. The simple version of the SI model assumes homogenous mixing of the population. Under this condition, at each time period, agents come into contact with another individual who is chosen uniformly at random from the population. If one agent is susceptible and the other infected, the infection is transmitted to the susceptible with probability β. These dynamics then repeat until all agents are infected or the simulation is stopped. While this process can be represented in a discrete formulation, it is most often presented continuously. The following set of differential equations describe the dynamics. A detailed discussion of continuous versus discrete simulation of such models can be found in Sterman et. al [39].

$$\frac{dS}{dt} = -\beta SI \tag{2.1}$$

$$\frac{dI}{dt} = \beta SI \tag{2.2}$$

$$\text{subject to} \;\; S + I = 1 \tag{2.3}$$

Solutions to this set of equations take the form of the classic S-shaped logistic growth equation,

$$I(t) = \frac{I_0 e^{\beta t}}{1 - I_0 e^{\beta} t}, \tag{2.4}$$

where I_0 is the fraction of the population initially infected.

In early stages, the infection spreads slowly as there are relatively few infected agents despite high numbers of susceptibles. The peak of new infections comes when there are roughly equal numbers of susceptible and infected agents. The number of newly infected individuals then decreases as the process saturates to the point where the entire population is contaminated.

The simple SI model always approaches an equilibrium where the entire population becomes infected. With no mechanism for an agent to recover or be removed from the population, a single infected individual will eventually contaminate everyone. However, patients routinely recover from an infection or are removed from a population due to immunity or, unfortunately, death. To account for these dynamics, many variants on the basic SI model exist. Below is a brief description of a few common extensions.

1. **SIR (Susceptible - Infected - Recovered):** Each period, an infected agent recovers or is removed from the population with probability λ. For a range of recover rates, it is possible that all infected individuals are removed before new susceptible individuals can be contaminated. This leads the epidemic to

die out. SIR systems are often characterized by the *reproduction number*, R_o. This number measures the ratio of the transmission rate to the recovery rate. If this ratio is greater than one, epidemics that infect all agents can appear in the system because new infections occur faster than old ones die out. For ratio's below 1, epidemics die out as individuals tend to recover before infecting others.

2. **SIS (Susceptible - Infected - Susceptible):** In this variant, agents can transition from being infected back to being susceptible. This does not allow for individuals to be removed from the population. SIS systems often oscillate between having large numbers of infected agents and large numbers of susceptible ones.

3. **SIRS (Susceptible - Infected - Recovered - Susceptible):** This variant adds a brief period of immunity to the model. Infected agents are removed from the susceptible pool for a short period of time, but eventually can be infected again. Diseases that mutate from year to year like flu viruses or for which vaccines lose effectiveness over time often display SIRS behavior.

4. **SEIR (Susceptible - Exposed - Infected - Recovered):** Often times individuals may not realize they are infected with a contagion. During this *exposure* period, they are still capable of infecting others, but because infection has not set in fully, they might not change their behavior.

The above is by no means an exhaustive list of possible combinations and permutations of states that might be added to compartment models. However, it is intended to show the flexibility of the approach. Solutions to more complicated compartment models often must be obtained numerically or by agent-based simulation. Finally, although a disease analogy was used to explain these models, the framework is general. Infection could be thought signing up for a service or buying a product.

For all its versatility, the simple SI model and its variants suffer from a few flaws. The assumption of homogenous mixing is unrealistic when considering human social systems. A given individual is not equally likely to come into contact with any other. People are far more likely to come into contact the close friend or family than acquaintances or strangers on the street. These basic approaches also ignore heterogeneity in the susceptibility of individuals in a population. Some people may have fewer anti-bodies for certain diseases or a higher propensity to adopter certain types of products. These features can dramatically affect the adoption process. Furthermore, the origin of the infection is a problem in SI models. They must be seeded with at least the tiniest fraction of infected individuals or else the system remains in the unstable equilibrium where the entire population is susceptible. Finally, there is no mechanism for individuals to be infected by external forces. In the case of technology adoption, external forces such as advertising or search might influence individuals just as strongly as friend.

2.2.2 The Bass Model

The Bass model is a close relative of the SI model [5]. Developed to forecast the adoption of a new technology, the Bass model provides a solution to the seed problem of the SI framework. Not only can individuals adopt a technology because a current adopter recommended it to them, they can also adopt due to some external force like media or search. Even if the Bass model is initialized so that there are no adopters in the population, adoption will eventually spontaneously occur. Moreover, it is possible to measure the relative efficacy of these two mechanism. Adoption occurring through other agents, *imitation*, is tracked along with adoption from external sources, *innovation*.

Using terminology from the SI model, susceptible individuals may become infected through transmission from another infected agent or through transmission from some outside source. In addition to the person-to-person transmission probability, β, there is an external innovation probability, α, of an individual becoming infected spontaneously. Also like the SI framework, the Bass model can be formulated as a set of differential equations.

$$\frac{dS}{dt} = -\beta SI - \alpha S \tag{2.5}$$

$$\frac{dI}{dt} = \beta SI + \alpha S \tag{2.6}$$

$$\text{subject to } \; S + I = 1 \tag{2.7}$$

The solution to the Bass model quantitatively and qualitatively follows the S-shaped curves characteristic of contagion spread.

$$I(t) = \frac{(\beta+\alpha)^2}{\alpha}\frac{e^{-(\beta+\alpha)t}}{(1+\frac{\beta}{\alpha}e^{-(\beta+\alpha)t})^2}, \tag{2.8}$$

When external sources are removed by setting $\alpha = 0$, the solution reduces to the logistic curve found with the SI model. Because the Bass model is typically used to predict life cycles of durable goods, recovery or removal states are generally not included.

The Bass model still suffers from the homogeneous mixing assumption despite including a mechanism for other influences beyond person-to-person. Furthermore, it is assumed that any outside influence from media or other sources is constant over the entire lifetime of the spreading process. This assumption is unrealistic considering huge media blitzes and trending popularity experienced by new technologies. Another important drawback of the Bass model is the difficulty of correctly estimating parameters early in a product's life cycle. Small errors in parameter values due to limited data can cause huge variations in predictions over long periods of time. By the time enough data has been gathered to make an accurate prediction, the adoption cycle is in the very latest stages making predictions less useful [5, 23]. Finally, as with the SI model, the Bass model assumes that all members of a population are identical with respect to their propensity to adopt a technology. A more realistic model should account for varying preferences or constraints. Certain individuals may buy every new gadget that is released, while others may wait until they can be sure of a product's quality. In their simplest forms, the Bass and SI models are silent on these issues.

2.3 Contagion on Networks

Whether solved analytically or simulated by discrete, agent based models, the traditional SI and Bass models' largest drawback is reliance on the homogenous mixing assumption. In most human systems, individuals are far more likely to come into repeated contact a limited set of others. Human social and contact networks are dominated by small groups of friends, family, and co-workers. These structures make a difference in the dynamics of contagion spread.

2.3.1 Small World Networks

Stanley Milgram's famous social search experiment was one of the earliest works demonstrating the importance of network structure in social phenomena. Milgram, a social psychologist, sent letters to a few hundred random participants throughout the American midwest with instructions to pass the letter to a target individual in Boston, Massachusetts. The instructions included basic information on the target such as name, occupation, and city he lived in (the exact address of the target was left out). If participants did not know the target personally (a common outcome in the early stages of the experiment), they were asked to pass the letter and the instructions on to someone that might. Before forwarding the letter, participants were asked to sign a roster listing everyone who had received it.

Milgram then waited to collect any of the letters that made their way to Boston. Amazingly, some fraction of the letters did reach their target destination. Those that did passed between an average of just six people, a much smaller number than might be expected [45]. Milgram showed that two randomly chosen individuals, separated by thousands of miles and having no immediate relations, can be connected by a very small chain of people. For this reason, these type of social search phenomena are now commonly referred to as *small world* experiments. Later, a mathematical definition of small world networks was developed. It states that the average length of the longest path, $\langle l \rangle$, between any two nodes grows logarithmically with the number of nodes in the network, $\langle l \rangle = O(\ln N)$.

Milgram also found that a large fraction of the letters were passed through a very small number of individuals. Moreover, these individuals seemed to have been chosen due to their occupation being similar to that of the target. This indicates that not all individuals in a network are equally important to the spreading process. Some have many connections, making them large contributors to infection, while others have few, increasing the chance that a contagion never spreads. The emergence of occupation as a common characteristic suggests that a real social networks have a degree of homophily. Homophily is the tendency for similar individuals, in this case those in the same occupation, to be friends. Colloquially, this property is captured by the phrase "birds of a feather flock together." The results from Milgram's experiment suggest that these facts are important to the way information is spread on a network and should be included in models attempting to model the it.

A few decades later, small world networks were famously formalized by Watts and Strogratz [49]. In their model, a network is created where individuals are represented as nodes and friendships by edges between the nodes. The small world networks presented by Watts and Strogratz begin as a ring network, where nodes are arranged in a circle and connected to k immediate neighbors. A fraction of all network edges are randomly rewired, creating a small number of long range edges cutting across the network. An example of a small world network can be found in Figure 2-1.

Watts and Strogratz found that this process produced networks similar to empirical measurements of social networks in two important ways. The first was that the maximum number of hops required to get from any node to any other (the network's *diameter*) grew slowly with the size of the network and decreased dramatically by adding just a few long range connections. The second property is known as the *clustering coefficient*. Colloquially, clustering is best described by the phrase "the friend of my friend is also my friend." In real social networks, if A is friends with B and B with C, then A is far more likely to be friends with C than is expected if connections were random. Watts and Strogatz's method for creating networks that have both

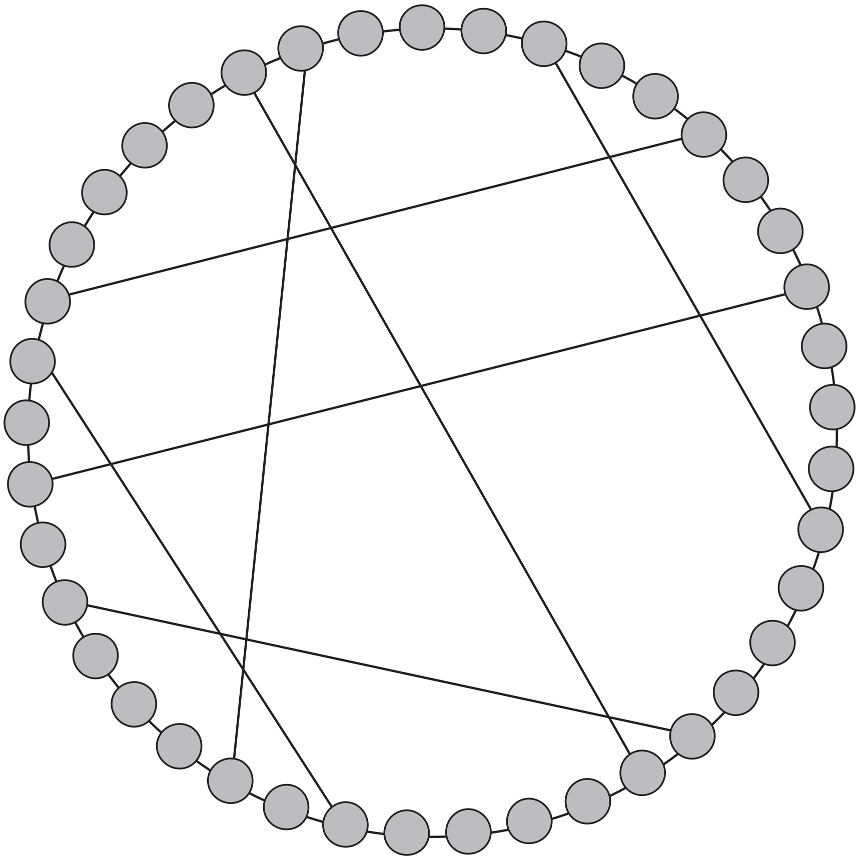

Figure 2-1: **An example of a Watts-Strogatz Small World network.** The Watts-Strogatz small world network is constructed by starting with a ring, $\langle k \rangle = 2$ shown, and randomly adding shortcuts between nodes.

small diameters and high clustering coefficients generated much interest in the area. The simulations and experiments performed with them have further demonstrated the importance of network structure. For example, replacing the homogenous mixing assumption of the SI or Bass models with a small world social network vastly increases epidemic spreading speeds. To analyze the properties of these more complicated systems, researchers turned to techniques for studying networks developed for use in statistical physics.

2.3.2 Physics of Networks

Statistical physics has provided many insights into the form and function of networks, though it wasn't until recently that models and techniques from physics were applied to networks in social systems. Tools for describing contagion on networks with arbitrary degree distributions[1] are particularly relevant in new social contexts. Using methods such as *site* or *bond percolation*[2], it is possible to explore the evolution of spread over a network while incorporating important features like those discovered by Milgram.

Percolation models traditionally described the flow of fluids through porous media like sand, but have been generalized explore the evolution of clusters in a network. When applied to contagion on networks, percolation models seek to answer the question: Will a disease outbreak become an epidemic and affect a large portion of the population? To answer this question, percolations models can be thought of as performing the following procedure either analytically or through simulation.

A network is generated by some procedure, for example, the small world process of Watts and Strogatz. In the case of *site* percolation, nodes begin in the 'off' state. A given fraction, θ, of nodes in the network are randomly switched the 'on' state. In the case of *bond* percolation, the same procedure is followed, but edges are turned 'on' and 'off' rather than nodes. After the fraction of nodes has been turned 'on',

[1] The degree distribution of a network refers to the distribution of the number of connections each node has to other nodes in the network.

[2] Site and bond percolation derive their names from models used to describe fluid occupying porous spaces in materials or flowing across channels between places

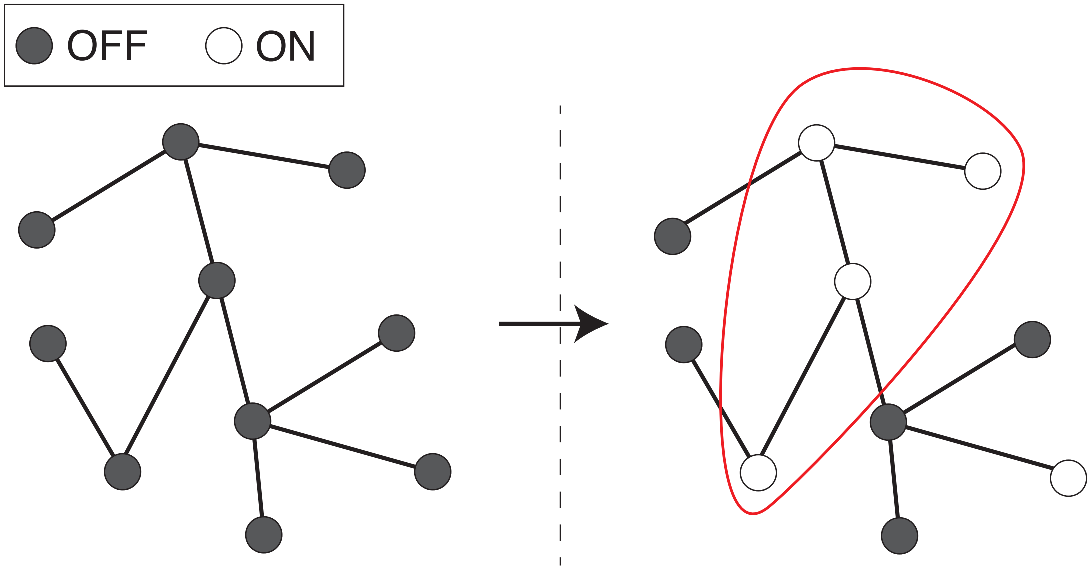

Figure 2-2: **An example of a site percolation.** The network on the left is initiated with all nodes in the 'off' state. A fraction of nodes are then turned 'on'. The largest cluster of 'on' nodes is then computed. In this case, a cluster of 4 nodes is created from the site percolation process. In the case of bond percolation, the same procedure is followed, but edges are turned off and on.

the largest set of nodes such that every node in that set can be reached from every other node is measured. This set is referred to as the *giant component*. The size of the giant component can be interpreted as the maximum size of the epidemic. It is common practice to measure the size of the giant component as a function of spreading properties like, θ [15, 35, 25]. Figure 2-2 depicts site percolation on a small network.

Because these models are stochastic, an ensemble of random networks is created for each value of θ. The ensemble average of the largest component size is calculated at that value. The procedure is repeated of a range of θ. The giant component size versus θ is then plotted. It is typical to find a critical value, θ_c, above which a giant component is formed, indicating an epidemic is present. For values of θ below the critical threshold, the contagion is trapped in isolated portions of the network and dies out before it can affect the whole population. It is generally stated that at this critical value, the system undergoes a phase transition to a regime that can support epidemics.

Percolation can be simulated while varying networks properties. One highly studied variant is a network's degree distribution. Networks with power law degree distributions that mirror human certain social networks, have much lower critical points than networks with more lattice like structures. This behavior reveals the importance of network structure to contagion spread. The very property of social networks that make social search so easy also ensures that disease spreads quickly.

The degree distribution is by no means the only property relevant to spreading. For example, *homophily* is often observed in networks where nodes have attributes, such as an individuals preferences. Homophily is the tendency for individuals to interact with others who are similar to themselves. People who enjoy a certain genre of music tend to be friends with those who share similar tastes [52, 24, 9, 51, 34]. This bias produces networks with community structures. A community may be a set of nodes that share many edges between members of that set, but have few connections to nodes outside it. This type of structure may help to localize certain contagions, but create hyper-important individuals. Those that connect two communities together serve as links between large portions of the network. These bridges, which are often *weak* ties like acquaintances, become very important in spreading information. Granovetter has described this phenomena as *the strength of weak ties* in his seminal work under that name [22].

The individual actors in networks have also been of frequent interest to research. Do nodes occupying certain positions, such as agents with many links, act as great innovators? Central actors may help diffuse a technology to the entire network, or inhibit information flow by refusing to adopt. Valente describes a number of systems, from doctors prescribing a new drug to farmers adopting a new seed. Each case tells a different story of who is important and what their position was in the network [46].

Massive popular interest in social networks has lead scholars to recognize the potential of using web applications to measure many of the characteristics described above. For example, it has been shown that different types of information follow different patterns as they are shared by millions of individuals on Twitter [32, 41].

Some information even takes on a life of its own, evolving into self-sustaining ‘memes’ [31]. In many cases, however, predicting the outcomes of such processes has proven extremely difficult [42]. More recently, studies have explored the many forces influencing the speed and success of information spreading such as blogs and traditional news outlets [54, 30, 41]. These studies have revealed a number of patterns whereby mass media drives conversation on social networks or vice versa.

While the vast majority of social network research measures properties of real world networks or the spread of simple contagion on them, there is another class of contagion that deserves mention. Whereas the single transmission of a disease requires two individuals, one infected and one susceptible, and is independent from other transmission in the network, there are many times that a more complicated mechanism is at work. For example, a person may decide to buy a new TV only after three of his or her five friends have done so. This of *complex contagion* has recently been explored by Centola. Surprisingly, Centola finds that spreading behavior of complex contagion is often the opposite of that found with simple contagions. Unlike disease, which spreads fastest in small world networks or networks with power law degree distributions, complex contagions spread faster on lattice networks. These results again demonstrate the subtlety and importance of considering network structure when modeling contagion[10, 11, 12].

Despite the power of network models to provide insights into contagion spread, they still suffer from a few deficiencies. While network structure does add an element of reality, nearly all of above models ignore how geography plays a role in network formation. Not only are people likely to be friends with a certain number of others, but those others are likely to be from locations near them. The computational complexities of metapopulations models have severely limited research in this area[15].

2.4 Binary Decision Models

Just as physics has contributed tools and techniques to the study of technology adoption and contagion on networks, the social sciences have also presented a number of methods. These approaches generally fall under the classification of *binary decision models*. Individuals in these models are faced with two choices, 0 or 1 [34]. In the context of economic decisions, a utility function is prescribed to the population and rational agents choose whichever option benefits them the most. This framework is especially useful for modeling the diffusion of contagion with significant externalities.

As López-Pintado and Watts note, simple mechanisms like those proposed in compartment models are often unable to differentiate between theories of how adoption happens. In short, there are many behaviors that can be encoded into agents to produce the same s-shaped adoption curves. Furthermore, these theories ignore strong externalities that may exist in the system, particularly with social goods. A person's valuation of a product or service can change dramatically as the number of adopters changes [34]. For example, a fax machine or Facebook account is useless unless someone else also has one. Similarly, competing standards must battle until one assumes enough market share that it becomes economically prohibitive not to comply.

The emergence of collective phenomena is also an important feature of these systems. Often times, global system behavior displays properties that are greater than the sum of actions taken by all individual actors in the system. Shelling's famous segregation model highlights the power simple behavioral models generate surprising emergent global phenomena. Seeking to explain the ways segregation might arise in neighborhoods, Shelling created artificial neighborhoods on checker boards. Agents of two races, 0 and 1, were distributed onto squares. Shelling then prescribed a number of utility functions for the agents, modeling how happy a person would be living in a neighborhood of a certain racial composition. If an agent is unhappy enough with their living situation, they move to a more suitable neighborhood provided one is available. As time moves forward, agents locate and re-locate themselves as the racial mix of neighborhoods change. The most shocking result of these models was that, even when tolerance, the preference of living in mixed race communities, was explicitly built into utility functions, segregation still occurred. Despite the attempts of individual agents to mix, the global system moved towards segregation[43]. Incorporating utility models with social influence allows for the explicit inclusion of preferences and externalities. Despite the difficulties in properly defining and interpreting utility functions, these techniques have yielded promising theories for tipping points and collective action problems. In these systems, the initial actions of a few can quickly snowball and move an entire system to one extreme or another [21, 2].

To borrow notation and definitions from López-Pintado and Watts, binary decision models generally begin by defining a population $N = \{1, \ldots, n\}$ of n individuals each with a set of possible actions $A = \{0, 1\}$. A vector describing the state of the system or the choices of all individuals then lives in the space A^n. Similarly, the vector of the actions for all others, excluding individual i, is defined as $\hat{a}_{-i} \in A^{n-1}$. A function $R_i : A^{n-1} \rightarrow [0, 1]$ is then defined to map the choices of all other individuals to a binary decision for individual i. $R_i(\hat{a}_{-i})$ is the probability that individual i choices action 1 given the choices of all other individuals. Adding another layer of realism, a set of weights, $\{w_{ij}\}_{N_{-i}}$, can be defined for the population. An individual may weight the action of close friend more heavily than acquaintance. The cumulative influence on i from all other individuals is then denoted by $k_i(\hat{a_{-i}}) = \sum_{j \in N_{-i}} w_{ij} a_j$. The social influence from all other individuals then maps to individual i's action by some function $r_i(k_i(\hat{a_{-i}}))$.

While this formalism seems a bit complicated, it is precisely in the mappings, r_i, and weights, w_{ij}, that externalities can be accounted for. For example, if $r_i(k_i)$ is an increasing function, it produces a behavior profile in which individual i abstains from adopting until a certain threshold influence, k_*, is reached. More complicated scenarios in which an individual only adopts when just enough others have adopted, but not too many. In these cases, an individual's decision to adopt is affected by his or her neighbors and that decision, in turn, affects others in the system.

To illustrate a model using the above framework, consider López-Pintado and Watts's description of a technology adoption model put forth by Shapiro and Katz [27]. Each individual has a utility function

$$u_i(a_i, k_i) = b_{a_i} - p_{a_i} + v_{a_i}(k_i), \tag{2.9}$$

where b_{a_i} and p_{a_i} are the benefit and price, respectively, to individual i from choosing action a_i and $v_{a_i}(k_i)$ is an externality dependent on the social influence from the actions of all other individuals. The sign of the externality is then estabilished by looking at how the difference in utility between choice $a_i = 0$ and $a_i = 1$, $\Delta u_i(\cdot, k_i)$ changes as social influence, k_i changes. Mathematically this can be written as

$$\frac{d\Delta u_i(\cdot, k_i)}{dk_i} = \frac{dv_1(k_i)}{dk_i} - \frac{dv_0(k_i)}{dk_i}. \tag{2.10}$$

If, as is commonly assumed in the case of technology adoption, individuals get more utility from adopting a technology when more people are also using, and less from abstaining when everyone is already using, this derivative is positive. In general, however, idiosyncratic arguments for the signs of each of these terms are specific to the particular system or phenomena being studied.

After the utility functions have been defined, the dynamics of a system can be studied. For example, in one of the simplest cases, the system is initialized so that each agent begins with a certain choice at the start. In each subsequent period, agents simultaneously update their actions based on their utility functions (which may depend on the actions of others). This amounts to a dynamical system where the state of the system now can be mapped to the state of the system in the next period. As is common with dynamical systems, fixed points where the system will reach equilibrium can then be identified.

Binary decision and collective action models provide a well defined notion of equilibrium. They offer a natural framework to study the stability and robustness of systems as well as what types of behaviors tip systems to more chaotic or emergent states. It is easy and intuitive to incorporate externalities, providing an unambiguous way to model interactions between individuals. However, these models often lack any acknowledgement of network structure and quickly become unwieldy when using anything but the simplest utility functions. Introducing any complexity in these functions requires numerical simulation.

2.5 Social Influence

Marketers and retailers are also very interested in understanding how information spreads as they try to increase sales and visibility for their companies. Recommendations are a large part of this process. In the era of social media, hyper influential personalities are increasingly important for advertisers and online stores can offer potential buyers recommendations based on the purchases of similar individuals.

Valente [46] describes a number of theories to explain how social influence may impact the diffusion of information. The reason that celebrities may hold so much influence is the enormous number of connections they share with others. The ability to transmit information to millions of people makes them powerful spreaders, capable of saturating a network with information very quickly. A network without highly connected people requires long chains of information passing to ensure everyone has heard the message. Moreover, others in the network may view well connected individuals as authorities (perhaps this is why they have so many connections to begin with). When a perceived authority adopts a technology or encourages others to do so, the message carries more weight.

On the other hand, there is considerable risk adopting a new innovation or recommending something to others. If the innovation fails, not only does the adopter lose his or her investment, but their reputation may also be damaged for recommending a poor product. With this in mind, well connected authority figures might be reluctant to adopt a new technology or pass information along to others. When this is the case, individuals on the periphery of the network, with few connections or power, become innovators. They simply have less to lose and are willing to shoulder the risks of innovation. To measure the plausibility of these theories in the real-world, researchers have studied online social networks like Facebook and Twitter. Influence can be quantified based on the ability of users to spread information through the system. The majority of these studies find that well connected individuals are modestly important, though perhaps not as important as one might think[37, 51, 3].

Furthermore, retailers have attempted to capitalize on the availability of high resolution sale data to extract purchasing patterns of customers and offering suggestions based on what similar customers bought. Similarly, items can also be grouped together based on these patterns so that stores can offer customers a list of complimentary goods. For example, the online retailer Amazon provides a list of items bought by similar customers as well as groups of items complimentary to the product being viewed. Entire business models have even been built around the idea of social purchasing. Groupon allows users to buy coupons to a restaurant or store and encourages purchasers to share the deal with their friends via social network sites.

However, there is a limit to the amount of social information people can process. In a 1992 article, British anthropologist Robin Dunbar estimated that a person could only maintain meaningful relationships with 100-200 others[16]. Dunbar did not have access to services like Facebook. Online social networking services not only store relationships, but provide near constant updates about those we may have lost touch with years ago. Moreover, they give individuals power to broadcast information to thousands of others instantaneously, all over the world. Still, can one person possibly keep up constant, meaningful relationships with thousands of people even with the aid of online tools? Recent research suggests the answer is ‘no.’ Dunbar’s number holds even for online social networks[20]. For the purposes of this thesis, this research can be used to make realistic estimates about the limits of social influence and network density in the diffusion of innovations.

In addition to recommendations people may receive from friends, there are other types of individuals who have influencing power. While celebrity endorsements have typically been a popular strategy of marketers, the ability of these celebrities to have direct contact with millions of fans via social networks has transformed their rolls considerably. For example, the *Colbert Bump* has been observed for political candidates who appear on the late night comedy show “The Colbert Report”. After an appearance, these candidates often find donations and polling numbers increasing by tens of percentage points[18]. Similarly, Oprah Winfrey’s book club is well known for turning average selling works into best sellers over night [38, 6].

Marketers and sales analysts have long sought to find patterns in customer purchase data. Better forecasts of sales in different cities or more accurate productions of how promotions or sales might affect the geographic shopping patterns of customers could help reduce costs associated with inventory stocks or staffing[19, 1]. These marketing and sales studies, however, are often held back by companies guarding valuable data from competitors and by the use of less sophisticated statistical techniques.

2.6 Including Geography

Geographic space is one of the most overlooked components of the diffusion process. Social networks are limited by peoples' ability to move and meet others. Geography, then, goes a long way in shaping social and contact networks, often giving rise to to strong spatial patterns among spreading phenomena. As the topology of social networks change, so do the patterns of diffusion. The introduction of high speed air travel along with the rise of instantaneous online communication has shifted the speed and cost of spreading information. This change can be seen most clearly by comparing the spatial diffusion patterns of the plague as it swept across medieval Europe during the 14th century, to more modern epidemic threats such as the H1N1 flu virus. Where as the former pandemic slowly marched from village to neighboring village at about the speed a wagon could be pulled, new flu viruses are delivered to major airports around the world in mere hours.

Because of the speed at which disease can travel on planes, great effort has been placed on developing optimal strategies for containing diseases before the reach pandemic stages. Airports must be quarantined, disrupted traffic must be re-routed, and steps must be retraced as quickly as possible to locate the source of a contagion and prevent disaster. To account for realistic travel and friendship patterns, attempts have been made to introduce space into models. *Metapopulation models* extend compartment models, allowing individuals to occupy states like susceptible or infected as well as a location in space such as a city, neighborhood, or town. Agents are then allowed to move between these locations. Mixing rates between metapopulations are estimated to simulate infection scenarios. Because generating and tracking multiple interactions between populations has traditionally been computationally intractable, most studies have remained aggregated to the city level, rarely modeling all individual level interactions at the same time [4].

Despite the difficulty in implementation, results suggest realistic geographies are very important to reproducing features of real contagion spread. Watts and Dodds [15] find that explicitly modeling interactions between places such as suburbs and city centers more accurately reproduces recurring epidemic patterns. In traditional models with homogenous mixing, simulated epidemics generally display smooth and predictable dynamics. A single quantity, the reproduction rate R_0, determines if an epidemic occurs and and how large it will be. The simulated epidemic is single peaked and affects a fixed percentage of the population before dying. In real disease data, however, multiple peaks are often observed. Metapopulation models, which assume homogenous mixing within each community (e.g. a small neighborhood), but allow individuals to move between locations (e.g. to different areas in a city) are able to replicate this observation. When disease is introduced into a single community, it creates a small epidemic among that group of highly connected individual. There is only a small probability, however, that the disease jumps to a neighboring community. If this does occur, another small epidemic is sparked. Metapopulation models accurately reproduce chain reactions of smaller epidemics where a community is infected and eventually. Furthermore, it is often due to the weak ties of Granovetter, the disease jumps to another susceptible community. The result at the global scale is a multipeaked epidemic.

In addition to disease spread, the diffusion of information is also influenced by geography. Farmers routinely rely on their neighbors for tips on which seeds are yielding the most bountiful and resilient crops. In a more mechanical sense, it is almost impossible to prevent seeds from blowing across property lines with the wind, leading to some transfer between neighbors. Neighbors are also often friends. Someone is far more likely to be friends a randomly selected individual who lives in his or her city than randomly selected person living in a similarly sized city on the other side of the globe. Do these local forces maintain their binding power in the face of new communications technologies that make it easier than ever to share information across great distances? A person can send an E-mail to a friend around the corner in the same time it would take to send that same E-mail to a colleague around the world (within a few milliseconds at least). As it turns out, geography even influences friendships online. Liben-Nowell et. al [33] studied a large online social network and found that users that the probability two friends were separated by a certain geographic distance was orders of magnitude higher for small distances than more large. More specifically, they found that the probability, p_r, of a friendship being separated by a distance, r, decreased as a power law with an exponent of roughly 1.2, but remained constant after a distance of 1000km ($p_r = r^{-1.2} + \nu$).

Further studies have suggested that the geographic structure of social networks may place constraints on spreading processes[36]. For all the apparent importance of geography, however, few studies have explicitly implemented geography into models. To date, there has not been a study devoted to understanding the balance of geographic bias in friendships and the flattening power of mass media and online communication technologies.

2.7 Gaps and Limitations

While the diffusion of innovations is by no means and understudied phenomena, there are a number of limitations and gaps in the above literature. Simple compartment models ignore the roles of social networks that were shown so important to real world situations. Binary decision models are capable of predicting equilibriums and incorporating externalities, but lack any notions of social structure. With few exceptions, all models, including those on networks, completely ignore geography. Empirical evidence indicates that individuals are not only connected to each other in a structured way, but they are also distributed in geographic space. Moreover, the above studies of innovation diffusion do not incorporate the mass media, either in the form of traditional media outlets or hyper influential celebrities.

The next chapter presents a model that address significant gaps in the above literature. It shows how the geographic distribution of individuals' differing propensities to adopt (such as early versus late adopters), combined with a preference for friendship with others who share similar tastes and geographic locations, are crucial features to accurately describe micro (at the city level) and macro (at the national level) adoption trends. Furthermore, the model includes an endogenous mass media agent that responds to adoption patterns of users as well as influences individuals to adopt an innovation. Based on adoption data from the popular social blogging platform, Twitter, the model of contagion to capture salient features. The next chapter is organized into three parts: (*i*) a presentation of spatiotemporal analysis of Twitter's (*ii*) a model simulating this adopting using insights from the case study to construct a network model, (*iii*) and finally results and discussion about important parameters and relationships.

Chapter 3

Modeling the adoption of innovations in the presence of geographic and media influences

3.1 Introduction

This chapter updates and unifies traditional models of information spread and technology adoption to more accurately reflect the novel economic, social, and geographic environments in which the spreading occurs. It expands on metapopulation models by embedding social networks in real geography to reflect the spatial distribution of social ties and better understand how local demographics and topology affect contagion. Furthermore, it introduces an endogenous media agent to a network model of information spread, capturing the role of hyper-influential social forces. The model is informed by a case study examining the *viral* (as it is colloquially referred) adoption of a social micro-blogging platform, Twitter, where the accumulation of users in cities across the country over a period of three years is quantified.

3.2 A Case Study of Twitter

As of December, 2011, the social micro-blogging platform, Twitter, had amassed roughly 300 million users globally. Started in San Francisco in early March, 2006, Twitter epitomizes the speed and efficiency with which an innovation is adopted by a population as well as its power to transform how we communicate.

3.2.1 How Twitter Works

As a web application, Twitter allows users to create a profile and generate short messages, or Tweets, of 140 characters or less. Upon creation, users are asked to choose a unique username and provide basic information about themselves such their current location and a personal description. This request is entirely optional and is not verified in any way, a feature that is taken advantage of by many false persona's on the site. Users can control who can see their Tweets by choosing to make their profile public or private.

A user's tweets appear on their main profile feed as well as on the feeds of those *following* them. Inversely, the tweets from users that a person *follows* will appear on the followers feed. There is an important difference between the *follower* and *followee* relationship that exists on Twitter and the *friend* relationship that is common on other social networks such as Facebook. Twitter allows for one way relationships. User A can follow user B and receive messages broadcasted by B, but if B is not also following A, B will not see A's messages. In this way, Twitter functions more as an information broadcasting and aggregation tool, where individuals can reach a large audience or receive updates from many others without the burden of maintaining a social relationship. It is common for celebrities to have hundreds of thousands, if not millions of followers.

This quirk in the usage and norms of Twitter creates interesting incentives and externalities for users to adopt. While Twitter was created as a social tool to disseminate information amongst friends, its information social structure lends itself to maintaining a large number of very weak, non-reciprocal relationships between hyperinfluencials like media and celebrities. This results in massive positive externalities, where Twitter becomes more and more valuable to users as more and more users accumulate on the site. Users are driven to the site by their friends, often people who live near them geographically, but are also joining to get updates from celebrity personalities halfway across the globe. This results in two opposing forces, one predicting strong geographic diffusions, the other disregarding geography entirely.

3.2.2 Twitter's History

Twitter itself was an experiment that grew from the company Odeo in March of 2006. Originally, the service was imagined as a group SMS messaging platform that would allow friends to quickly and easily communicate their activities online and through mobile phones. Twitter remained a relatively small operation until South By Southwest (SXSW), a popular annual tech and music conference, in 2007. The company demoed its technology by coordinating realtime feedback on the event. Twitter was a hit, but its user base remained relatively small, confined to young, tech-savvy demographics (statistics provided in later sections)[1]. During the first two years of existence, Twitter did not engage in traditional media advertising. Instead, it relied almost entirely on word-of-mouth buzz. Twitter never participated in a major advertising campaign on traditional media such as TV or radio and even refrained from developing any real business model until the later stages of its growth.

By early 2009, Twitter had amassed millions of users. Around this time, celebrities began to realize the power of the platform to connect with fans. Actor Ashton Kutcher embarked on a campaign to be the first Twitter user to reach 1 million followers (people subscribed to his feed). On April 17th, Mr. Kutcher appeared on the Oprah Winfrey Show to announce he had succeeded in his goal. On the same show, Oprah herself signed up for Twitter and encouraged viewers to do the same. Mr. Kutcher's campaign, followed by Oprah's endorsement generated a huge increase in Twitter users as well as traditional media buzz for the site. Later in 2009, Twitter again found itself in the news. This time, the world was debating the

[1] `http://www.nytimes.com/2010/10/31/technology/31ev.html`

roll of social media in coordinating protests in the Middle East, most notably Iran. As a symbolic acknowledgment of Twitter's influence on the world, Collins English dictionary officially added "Twitter" as a verb at the end of the year[2]. Since the end of 2009 Twitter's user base has grown to nearly 300 million worldwide. It has become a platform for pop-culture and communication, sparking countless other businesses and scholarly work. Moreover, the availability of an open Application Programming Interface (API), has allowed developers and academics to download and analyze data from millions of users and billions of Tweets, making it an incredibly rich source of data.

3.2.3 The Dataset and Descriptive Statistics

To understand the adoption of Twitter in both space and time as well as the role of media, this chapter analyzes data on when and where Twitter users in the United States signed up for the service. Data were collected by Cha et. al [13] in August of 2009. Cha obtained permission from Twitter to gain access to and copy information on the roughly 55 million US users signed up at that time. In many cases, however, a person may sign up for an account and never use it or make multiple dummy profiles. To account for this, Cha et. al removed users based on the level of activity they generated. This left nearly 3.5 million 'active users' in the US. For each active account, the time and city of creation was recorded. Because researchers were given direct access to Twitter's servers, data could be collected starting immediately after Twitter's launch in late March, 2006 and ending after its first massive surge in popularity in late August, 2009. In total, users signed up in roughly 16,000 unique cities across the country. The data is restricted, however, to cities where at least 1000 users had signed up over the 3 years to ensure sufficient statistical power. This thresholding left 408 cities to work with and includes roughly 2.3 of the 3.5 million, or 66%, of all active users. For the remainder of this chapter, analysis is restricted to this thresholded data set.

Beyond word-of-mouth recommendations from social contacts, individuals can

[2] `http://mashable.com/2009/07/06/twitter-in-the-dictionary/`

also learn about Twitter through search or more traditional media outlets. Google's Trends and Insights web application is used to incorporate these mechanisms into the analysis. This application allows users to obtain time series on the search and news popularity of keywords and terms in Google's extensive database. For this study, weekly search and news reference volume were obtained for the period covered by adoption data (March, 2006 - August, 2009). Though Google is certainly not the search engine on the internet, it is the most popular. It is also unlikely that search behavior differs significantly between major search engines. The number of people searching for "Twitter" on Google should be representative of the number of people searching for "Twitter" elsewhere on the internet. Moreover, Google's popularity makes it very attractive for media outlets that want to generate as much traffic to their stories as possible. This suggests that Google's index of news stories about a topic is comprehensive and accurately reflects major patterns of media buzz. For these reasons, the number of people searching for the term "Twitter" on Google and the number of articles referencing "Twitter" in Google's news index provide excellent proxies for individual search behavior and mass media volume.

Google only allows access to normalized search and news volume. The maximum over a given period is set to a value of 100 and all other values are scaled to preserve relative magnitudes. This prevents one from knowing the absolute number of times a term was searched, but its sufficient to track relative popularity of terms and dynamics in time. Though the feature is not used in this analyses, it is also possible to break down search and news volume by geography in addition to time. Again, absolute numbers are normalized against the maximum volume in a particular region, but it is still possible to compare the relative popularity of a search time over time and space[3]. For the purposes of this research, however, these data are used understand the dynamics of mass media on an aggregate, nationwide scale.

Fig. 3-1 displays national Twitter adoption trends as well as search and news volume. The cumulative number of users qualitatively matches the classic S-shaped

[3] More information on the precise scaling and normalizing of these data can be found at `http://www.google.com/intl/en/trends/about.html`.

adoption curve found in many innovation diffusion contexts. Adoption begins slowly as few people know about an innovation or are hesitant to adopt. As the innovation gains traction, adoption takes off and a majority of the population adopts at a rapid pace. Eventually, this pace slows as it becomes too difficult to find individuals who have not yet adopted. Having reached all potential adopters, the process saturates and eventually dies out. National Twitter data distinctly show the first two stages and suggest a slow down into the third stage. It should be noted, though, that Twitter's user base continued to grow at very high levels after the end of this studies data collection period. Unfortunately, this study is limited by data collected. It assumes that the slowdown happening at the end of 2009 represents saturation of a particular market. Additional, independent growth periods may occur later.

While cumulative trends show traditional patterns of adoption, week-to-week growth reveals more interesting dynamics. In contrast to traditional S-shaped adoption curves that displaying a smooth increase and decreases in the number of new users per week, Twitter data is highly variable. Very rapid increases result in huge spikes of users signing up one week with relatively few doing so in the next. Moreover, these spikes correlate very closely with spikes in both search and news volume. This suggests that a strong relationship exists between media and adoption and that our use of Google news and search data is appropriate. In addition to the strong correlation, Google data indicate that media coverage of Twitter was nearly non-existent during the first two years. Over this period, Google search volume was highly correlated with user growth, suggesting that individuals, having heard about Twitter through a friend, went searching for the web application on Twitter and that many ended up signing up.

During later periods, the spikes in media coverage and adoption suggests some important discrete events occurred. More careful analysis of reporting during the weeks surrounding spikes reveals major news events like celebrity endorsements and political unrest. In early 2009, actor Ashton Kutcher began a campaign to become the first Twitter user with 1 million followers. To announce success in reaching his goal, Mr. Kutcher appears on The Oprah Winfrey show on April 17th, 2009. During the show

Oprah herself signed up for Twitter and urged viewers to do the same. This endorsement resulted in the single largest weekly increase in Twitter users over the period studied here. In the weeks following the show, this adoption rate quick fell back to characteristic levels. In the fall of 2009, adoption rates spiked once again. This time, rather than a celebrity endorsement, Twitter was in the news for its roll in coordinating political unrest throughout the Middle East. Specifically, news correspondents and bloggers were debating the roll of social media in the protests occurring in Iran. The corresponding spike in new Twitter users suggest that increased media exposure encouraged individuals to sign up. This event highlights the strong endogeneity that exists between growth rates and media attention. Data show that the media responds to the adoption it produces. This is much different than the traditional modeling of media [5, 26]. A powerful media agent that both grows with adoption and experiences random shocks is added to the model.

Cumulative adoption data can be used to study characteristics of individual users. For example, some individuals have much higher propensities to adopt a product in its early stages. These early adopters may have the most to gain from adopting or may simply be the type of people who must have the latest gadgets. Conversely, some individuals are hesitant to adopt a technology and the associated costs. They lag behind the rest of the population to make sure of quality. Placed in the context of a social network, heterogenous populations matter. If an early adopter is only friends with laggards, they will never be the first to learn about a technology because their friends will not care to tell them. They will be isolated from innovation.

Procedures from the diffusion of innovations literature are followed to measure the prevalence of these types in data. Adopters are labeled according to where their adoption times fall relative to the distribution of all other times. Those who adopt greater than 1σ (standard deviation) before the average adoption time are labeled as early adopters. Those adopting less than 1σ before the mean adoption time are the early majority. Late majority and laggards are labeled by similar intervals after the mean time. For more on the motivations behind this, see Rogers [40]. Time series for different cities around the country reveal effects on adoption heterogenous types

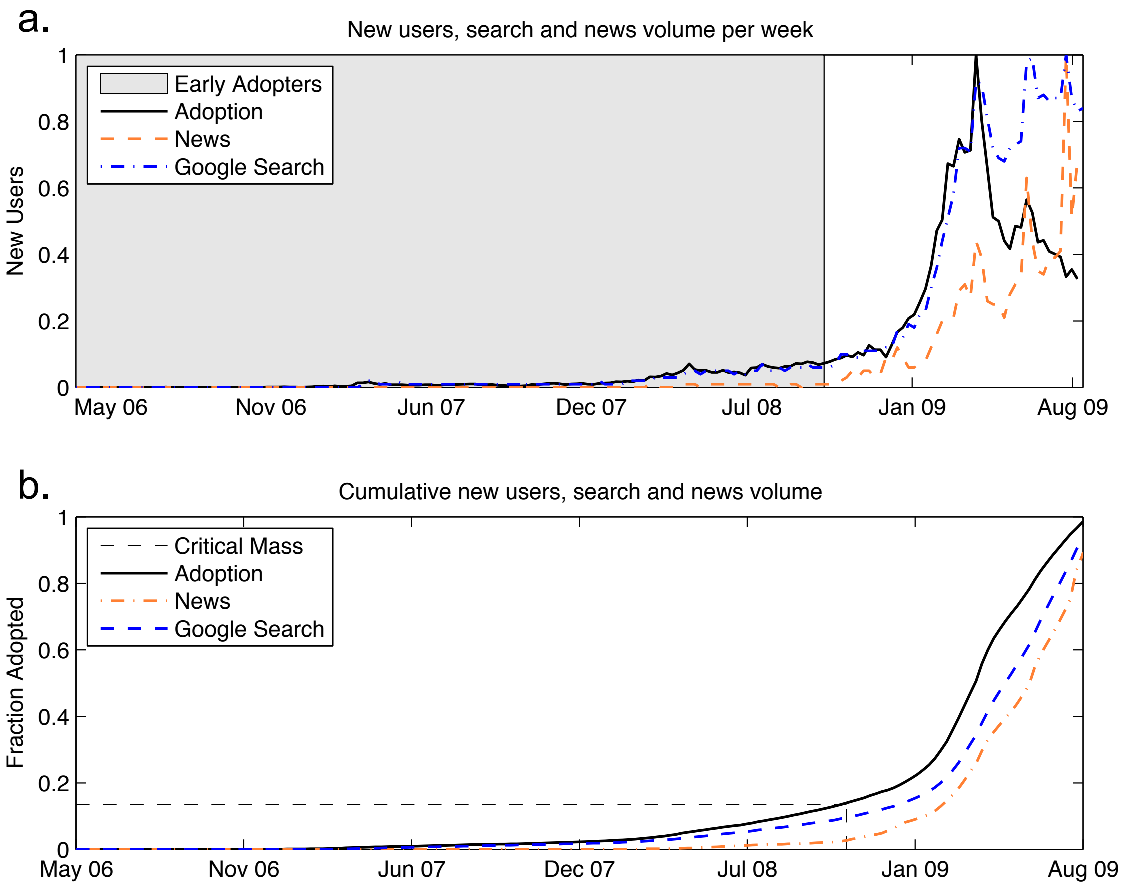

Figure 3-1: **Plots of weekly national adoption.** (a.) The number of new U.S. Twitter users is plotted for each week, normalized by the maximum weekly increase during the entire period of data collection. (b.) The cumulative total number of U.S. Twitter users is plotted for for the same time period. Google search and news volumes are normalized such that the maximum value is 1.

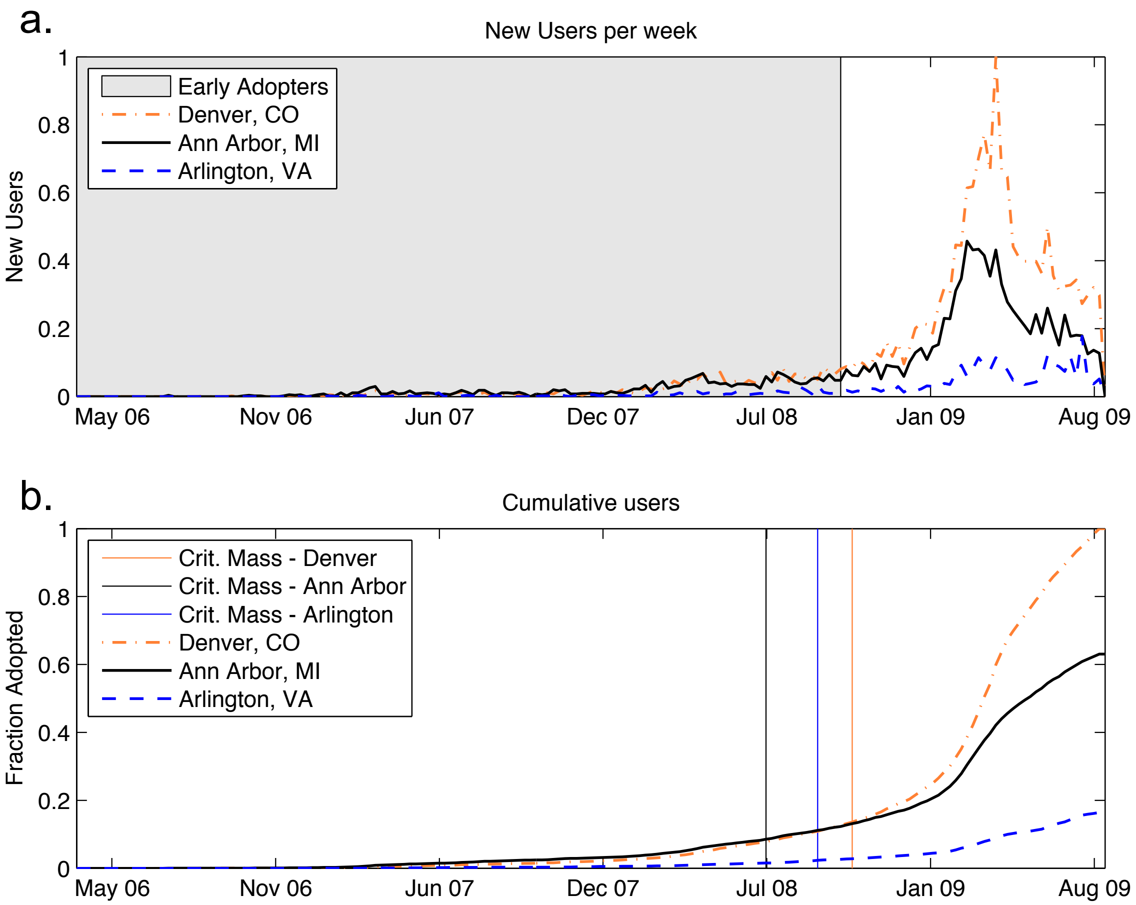

Figure 3-2: **Plots of weekly adoption for select cities.** (a.) Time series display the number of new U.S. Twitter users for three separate locations (Ann Arbor, MI, Denver, CO, and Arlington, VA) from mid-March 2006 through late-August 2009, normalized by the largest weekly increase in Denver users. (b.) Shows a plot of the cumulative fraction of each city's user base normalized by the total number of users in Denver, CO.

can have in both space and time. Fig. 3-2 shows three separate locations across the country representing a young, early adopting demographic (Ann Arbor, MI), a large metropolitan consisting mostly of late majority adopters (Denver, CO), and a mixed area (Arlington, VA).

The labeling allows the composition of each city to be measured in terms of the percentage of users that are early adopters, early majority, late majority, or laggards. This analysis also serves to normalize locations with respect to population. Large cities will have more early adopters than small towns, but as a percentage of total population early adopters may be scarce. Qualitatively, analysis suggests that cities with the largest fractions of early adopters tend to have large universities or are technology centers. These institutions attract large numbers of young, tech-savvy persons, just the type that are likely to adopt social web applications. Later, numerical simulations show that the empirical composition of cities and the demographics they represent are critical to reproducing spatiotemporal diffusion patterns.

The key moment in the diffusion of an innovation comes in the transition between slow initial growth and rapid adoption. Colloquially, this moment is known as achieving *critical mass.* Again following conventions from the diffusion of innovations literature, critical mass is defined as the point when an innovation is adopted by 13.5% of a population. For this study in particular, a city is said to have reached critical mass if 13.5% of all eventual users in that city have signed up [46]. Ideally, the population through which Twitter is diffusing in consists of all persons with internet access in a city. Unfortunately, data on this population is unavailable. Again assuming that the adoption process ends after August 2009, the total number of users at this time is used as a proxy. The timing of the media's involvement adds more confidence to this definition. Google news volume shows almost no media coverage for the first two years of Twitter's existence. However, news volume picks up just as Twitter reaches critical mass nationally. While it is impossible to determine if this relationship is causal, it suggests that the tipping point is meaningfully operationalized by the stated definition of critical mass achievement.

The remainder of this chapter presents an explanation and prediction of spatiotemporal patterns of critical mass achievement. Fig. 3-3 shows a series of snapshots in time indicating when various cities reach critical mass. These snapshots reveal the diffusion path of Twitter from its birthplace in Silicon Valley, to college towns such as Cambridge, MA, Ann Arbor, MI, or Austin, TX, to metropolitan areas such as Los Angeles, CA, or Denver, CO, then finally to more suburban and rural areas.

Just as individuals users were labeled as an early adopter or a laggard, cities were also placed into groups according to when they reached critical mass relative to the entire population. Table B.1 in Appendix B presents a complete list of cities

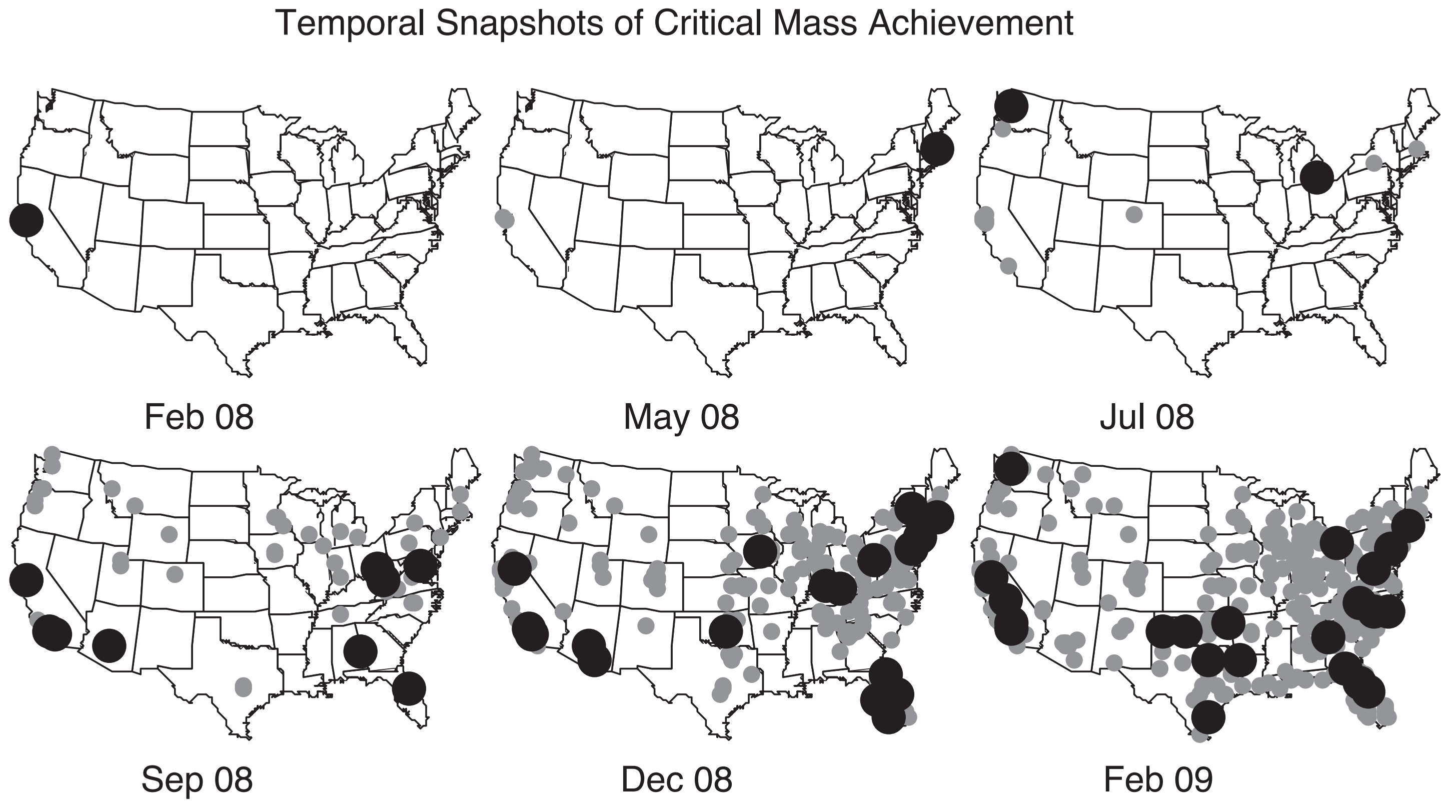

Figure 3-3: **Temporal snapshots of critical mass achievement at locations across the US.** For snapshot, the smaller, gray markers indicate locations that have already reached critical mass. The larger, black markers denote locations that achieved critical mass during that week. We note that locations achieving critical mass at very early times are clustered around Twitter's birthplace, San Francisco, CA, suggesting local word-of-mouth diffusion. There are, however, a few locations on the other side of the country, namely the suburbs of Boston, MA that are equally early in adoption, contrasting local diffusion with the flattening effects of the Internet.

and their classification. A qualitative assessment of these groups reveals the type of demographic information that can be inferred from looking at the adoption of web applications. Nearly half of the cities labeled as "early adopting" are home to major colleges and universities. Large proportions of their populations are younger, tech-savvy students. Conversely, laggard cities tend to be in more suburban and rural areas with very different demographics. Major cities, which house a diverse set of inhabitants, fall in the middle.

To summarize, descriptive statistics of the data set reveal features of Twitter's adoption in the US. While national cumulative adoption qualitatively follows a traditional S-shaped curve, analysis of week-to-week growth indicates a more variable process. Twitter grew relatively slowly for the first two years of its existence and did not generate media interest. During this time, Google search volume is highly correlated with adoption. After Twitter achieved critical mass nationally, the search volume decoupled from adoption. News coverage increased due to discrete news events and adoption rates increased dramatically. The composition of cities was measured in the fraction of each population measured to be early adopters, early majority, etc. Moreover, local critical mass achievement times were measured for each city. Cities with younger, tech-savvy populations reached this tipping point sooner, while suburban and rural communities lagged behind.

The remaining sections of this chapter present a model capable of replicating the empirical results described above. Though the model is general, it is tested based on its ability to replicate the dynamics of Twitter's adoption. Incorporating important features such as city composition and an endogenous media, the model simulates the diffusion of an innovation in a group of a cities. Critical mass achievement times as well as total users at the end of the simulation are benchmarks for performance.

3.3 Model

To capture both geographic effects as well as media influence, the following model is created:

(i) The first step initializes the agent population and social network. Innovation diffusion is simulated by a mechanism resembling the susceptible - infected (SI) model. The SI model is also a special case of the Bass model that is widely used in the diffusion of innovations literature. A population of N agents is created and each is placed into one of L cities. This creates a set of city level meta-populations and introduces geography into the model. Each agent can be one of two types, an *early* or *regular* adopter[4]. The geographic placement and agent types can be chosen to reflect empirical measurements of Twitter data. Furthermore, agents can be distributed in space to reflect measured populations in cities. The composition of these cities in terms of agent type is also controlled. For the purpose of calibrating and validating the model with real Twitter data, if a city was measured to have 4% of all US Twitter users, 4% of our agents are placed there. Of the agents placed in that city, if the composition was measured empirically to be 30% early adopters, 30% of agents will have an early adopter type. The remaining are marked as regular.

A social network is formed by connecting agents with links. Links can be assigned randomly to replicate the homogeneous mixing assumption of most compartment models. It is also possible to connect agents according to empirical characteristics observed in on-line social networks. For example, Liben-Nowell et al. [33] show that p_r, the probability of being connected to someone located a distance r from your city, follows a truncated power-law, $p_r = r^{-\gamma} + \nu$, where $\gamma = 1.2$ and ν is set such that the probability of connection becomes roughly constant for distances greater than 1000km. It is also possible to set other network properties such as degree distribution and density to reflect different topologies.

(ii) Next, dynamics are added to the simulation. At any given time, an agent can be in one of two states, susceptible (S) or infected (I). A small fraction of agents are initialized as infected to seed the contagion. Spreading is modeled over a series of T time periods, where the number of agents in each state is tracked (subject to $S(t) + I(t) = N$). Each time period, all infected agents attempt to infect their

[4] To simplify the model, all users who are not considered early (early majority, late majority, and laggards) are considered to be a regular adopter—.

neighbors. With probabilities β_r and β_e, a regular or early adopter, respectively, will heed a recommendation and adopt the technology. The ratio, $R = \frac{\beta_e}{\beta_r}$ controls differences in propensity to adopt for early versus regular adopters.

These features mimic social dynamics that suggest the pressure to adopt increases as more friends adopt[46]. Some models assume that an individual will adopt an innovation once a specific number [22, 21, 50] or proportion [10] of their contacts have also adopted. Others have found evidence that occupying similar positions in social networks is more predictive of adoption [7]. While this model does not attempt to test these hypotheses, Kleinberg has suggested that the dynamics of these adoption schemes are quantitatively similar [28].

(*iii*) In addition to word-of-mouth spreading, a media agent is also included. This agent can be thought of as an influence in addition of word-of-mouth spreading. Each time period, the media broadcasts its message to adopt a technology. Having heard the message, each agent flips a coin determining if adoption occurs. The media transmission probability is given by, $\Pr(media\ infection) = \alpha M$, where $\alpha \in [0, 1]$ is a model parameter, and M is the endogenous media volume. Media volume itself is determined as a function of the number of previously infected agents, $I(t-1)$, and a random term ϵ such that $M(t) = I(t-1)^\gamma + \epsilon$. For convenience, media volume is normalized so that, $M(t) \in [0, 1]$. The parameter, γ, reflects the super-linear growth displayed in Google news media volume. Finally, the size of random shocks ϵ is set on the order of $M(t)$, reflecting stylized features seen in Google News volume data.

In essence, the amount of media exposure an innovation is given depends explicitly on the number of people who have adopted as well as a random error term. Just because the media is reporting on a new product, however, does not mean a consumer will adopt it. To model this, the parameter α is included. This adjusts how receptive agents are to the media. The probability that any given agent will adopt due to the medias influence, αM, is interpreted as the product of how much the media is reporting and how closely an individual is listening.

The model was implemented in Python utilizing the open source SciPy and NumPy libraries to perform calculations and statistics. A full description of the implementation, including model parameters and input-output functions can be found in Appendix C.

3.4 Results and Discussion

3.4.1 Replicating standard SI model

The first results validate the model for known parameter regimes. Parameters are set so that the simulation reduces to the traditional SI model. Only one type of agent is modeled (setting $\beta_r = \beta_e$) and the media is turned off ($\alpha = 0$). Each of the $L = 408$ cities are populated uniformly with 1000 agents for a total population of $N = 408,000$. The network is then initialized to have a completely random spatial distribution of links and a Poisson degree distribution. A Poisson degree distribution is chosen because the structure of the adoption network is more selective than a scale free structure found in measurements of all connections in online social networks [30, 53, 20]. For example, Leskovec et al. [30] found that individuals who recommended a product to tens or even hundreds of contacts influenced no more purchases on average than those who sent recommendations to just a few friends. Thus, the expected number of people who can influence a person to adopt a technology is smaller than the number of acquaintances they have and the distribution is not likely to be long tailed. Scaling these numbers to fit the simulation size, a reasonable average degree of $\langle k \rangle = 7$ is chosen.

Fig. 3-4 displays the simulated number of adopters per week for a variety of values for β. The simulation was run 500 times for each of the parameter settings. Bands surrounding the average represent ranges between which 75% and 95% of simulations fell. In this simple form of the model, it is not possible to reproduce the empirical shape of the cumulative adoption curve seen in the Twitter case study.

Next, geography is added to the model. City populations, spatially embedded friendships, and early adopting agents are introduced. Agent types are also de-coupled by assigning different propensities to adopt. The best results were obtained by setting

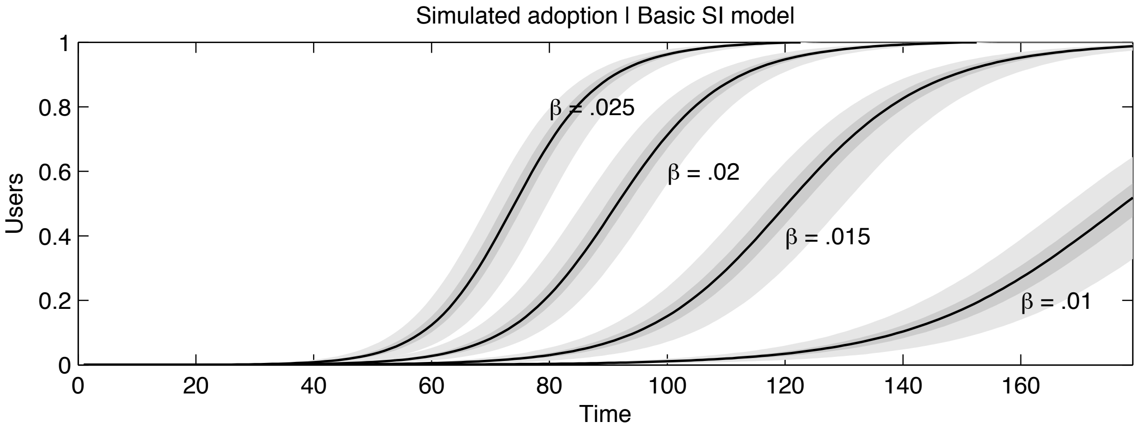

Figure 3-4: **Verification of the basic SI model.** Four different transmission rates β are displayed, each run 500 times and averaged. The bands surrounding the average value are bounds containing 75%, and 95% of simulation runs.

early adopters to be three times more likely to adopt than regular adopters ($R = \frac{\beta_e}{\beta_r} = 3$). In addition to heterogenous agents, the spatial properties of the network are also changed the topology on which adoption occurs. Now, early adopters are also concentrated in specific locations and are far more likely to be friends with the people around them. This type of heterogeneity affects the size and growth of large clusters of agents that are all connected to each other. More generally, a cluster is a set of nodes for which any node in that set can be reached from any other node in the set by following links only between nodes in the cluster. As defined in Section 2.3.2, if the largest cluster in a network contains a significant fraction of all nodes in a network, it is referred to as the *giant component*. The size (number of nodes) of the giant component is an important quantity that greatly affects spreading processes on networks. For example, in traditional network disease epidemic models, the size of the largest cluster in the network is an approximation of an epidemic's maximum size.

In the diffusion model presented here, the most important cluster is the giant component of early adopting nodes. This is the largest set of early adopting nodes such that each early adopter in the set can be reached from any other early adopter also in the set by hopping between early adopters in that cluster. A small giant component of early adopters indicates that most early types are connected to regular adopters. These regular adopters effectively isolate the early adopters, diminishing the chance they will learn about an innovation. For example, if an early adopter is only friends with regular adopters, they must wait until a regular type adopts a technology before they even learn about. This effectively immunizes the early adopter. On the other hand, if an early adopter is friends with other early adopters, the chances they will adopt innovations at an early time is greatly increased. The innovation can then spread very quickly through the population of early adopting types because they have access to information and are willing to adopt.

Results from simulations show that homophily affects the size of the giant component of early adopters and this, in turn, affects the diffusion process. In general, an increase in homophily increases the size of the giant component of early adopters. Interestingly, however, homophily based solely on agent type (i.e. early versus late adopter) is not enough to reproduce the observed trends in the spatiotemporal diffusion of information. A more subtle type homophily must be present to ensure that the early adopters are connected to each other. Homophily due to agents' tendencies to be friends with nodes who are close spatially is also required. To introduce the latter type of homophily, two types of spatial networks are simulated, homogenous mixing and spatially embedded networks. The fraction of similarly typed neighbors that each agent prefers (traditional homophily) is also varied. Simulations suggest that a giant component of early adopters is formed at much lower levels of network homophily when in spatially embedded networks. In other words, spatial social networks tend to have much larger giant components of early adopters for a given level of network homophily.

The intuition for this result is as follows. Homophily by type ensures that early adopters will be friends with other early adopters, creating clusters within the network. For reasonable levels of homophily by type, however, these clusters are not connected to each other because the density of early adopters is too low. This prevents a giant component from forming. However, if early adopters are more likely to be connected to other early adopters who live near them, all the early adopters in a particular city will be connected in a cluster. Now, a single connection between an

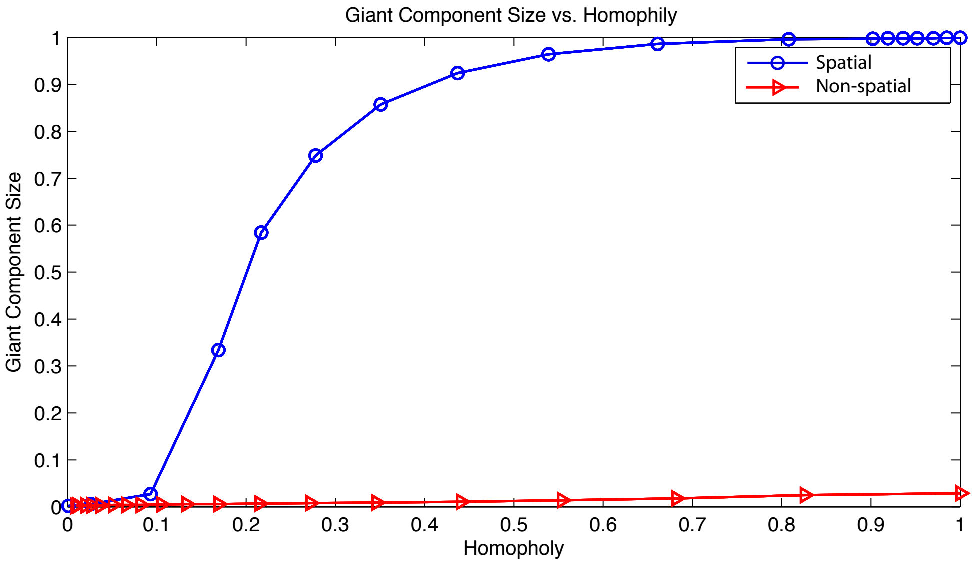

Figure 3-5: **The size of the giant component plotted against homophily.** Two configurations are shown, one in which the social network is explicitly spatial, the other ignoring geography of nodes. The figure illustrates that preference for friendship with similar agents is not enough to connect early adopters in a giant component and that spatial friendships are produce this structure.

early adopter in one city and an early adopter in another effectively connects all the early adopters in both cities in a larger cluster. This makes it much easier for a giant component to form.

Fig. 3-5 plots the size of the giant component of early adopters produced at a given level of homophily measured among early adopters for networks either spatially embedded or not. Here, homophily is defined as the average fraction of an early adopter's friends who are also early adopters. These estimates were obtained by creating and consolidating results over 100 networks, each with $N = 10,000$ nodes and a given level of homophily, then measuring the size of the giant component. For the remainder of this chapter, all configurations labeled *spatial network* can be assumed to have a giant component containing over 95% of all early adopters.

To explore the ways in which giant components of early adopters affect adoption, Fig.3-6 compares the predicted and actual times of critical mass achievement when diffusion is simulation on spatial versus non-spatial networks. In the absence of a giant component, nearly all cities peak at the same time. Simply placing different

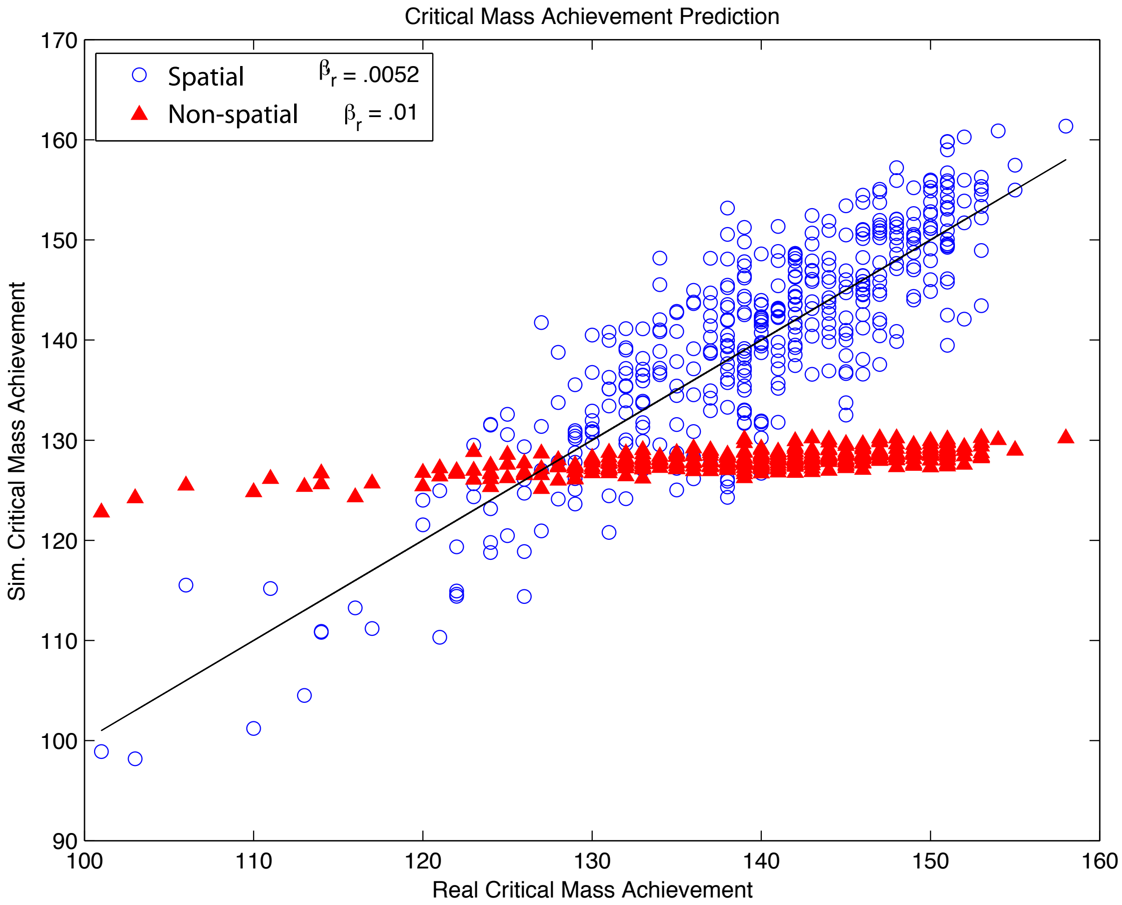

Figure 3-6: **Simulated critical mass achievement times are compared to times measured from Twitter data.** We find spatially embedded friendships are necessary to reproduce the inter-city spread of Twitter.

numbers of early adopters in cities is not enough to change diffusion patterns. When spatially embedded friendships are introduced such that a giant component of early adopters is formed, city-to-city patterns are recovered. Though global cumulative adoption can be reproduced without the spatial social network, adoption cannot be geographically resolved to the city level. Embedding the social network in real space accurately predicts critical mass achievement times in most cities. Fig. 3-7 shows box plots of simulated times compared to empirical data for selected cities. Cities have been divided into four groups based on when they reached critical mass relative to all locations.

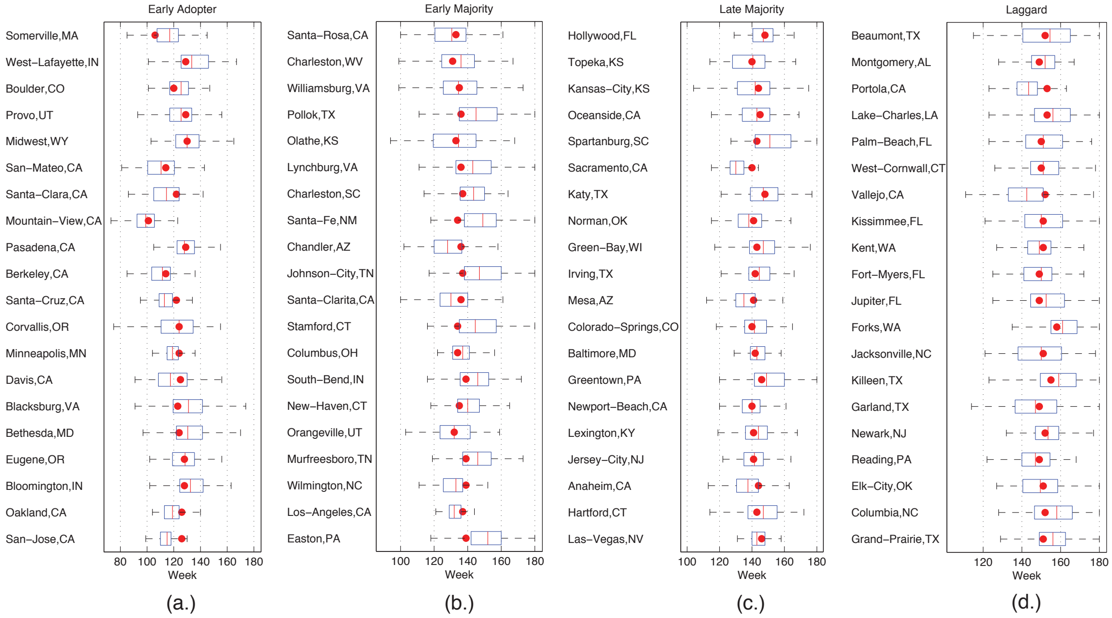

Figure 3-7: **Simulation results are compared to actual critical mass achievement times for different subsets of locations.** Borrowing from the diffusion of innovations literature, we use four groups (a.) Early adopting, (b.) Early Majority, (c.) Late Majority, (d.) Laggards. We are able to reliably predict adoption times for cities in each category.

3.4.2 Media Influence

While spatial social networks accurately predict critical mass achievement times for innovation diffusion on a city level, comparing simulated to real adoption trends reveals discrepancies at later times. Fig. 3-8 compares predictions of national adoption with the model conditions from the previous section. Simulations start diverging from reality around week 120 after launch, indicating key features are missing from the model. Moreover, divergence begins around the time Twitter reached critical mass nationally. Up until that point, very little media coverage was present. After critical mass is achieved, media volume begins to increase substantially. This transition can be used to measure the relative strength of word of mouth spreading compared to media influence.

Predicting when individual media events like celebrity endorsements will occur is beyond the scope of this work. However, adoption in the presence of media can be simulated with empirical data on news volume. Exogenous media volume from Google News data is input into the model for $M(t)$ to fit parameter values for the propensity to listen to media, α. In order to achieve the national adoption pattern similar to that seen in real data, agents must be highly susceptible to media influence. Parameter values of $\alpha \approx 0.15$ are required to accurately reproduce adoption trends. Comparing simulations with and without mass media suggests that its presence was responsible for for at least half of the Twitter's user base. Most of these users adopted in later stages when media volume was very high. Coupled with early results showing the importance of homophily and geography during the early stages of spread, the model presented in this chapter paints a much more complete picture of adoption than traditional approaches. Both aggregate and local trends in space and time can now be simulated and predicted.

The disadvantage of the above procedure is the exogeneity of media influence. Data on news volume must be known in advance in order to predict adoption. To solve this problem, the model is extended to treat news volume as endogenous. Endogenous mass media is implemented as described previously as step *iii.* of the model introduction in Section 3.3. Reflecting trends seen in the real data, the growth of media volume is super-linear with respect to adopters and random spikes in media coverage are introduced to reflect discrete and unpredictable media events. Numerical simulation shows an exponent of media growth with respect to adopters,$\gamma = 3$, produced reasonable fits to real data.

Fig. 3-8 displays simulation results for various model settings described in this paper. While spatial friendship networks are able to reproduce early adoption trends, real data quickly diverges in later times. Introducing an endogenous mass media agent which grows super-linearly in the number of current adopters as well as random media spikes, produces much more accurate adoption trends and reflects features seen real media coverage.

In light of a globalized world with near universal access to the Internet, previous models of adoption fail to characterize the interplay of media and word of mouth. During early stages, when spreading occurs primarily through word-of-mouth, simulations show that adoption is strongly correlated with traditional demographic covariates. Early adopting cities tend to be those with large, young, and tech-savvy populations. Moreover, these demographic groups must display high levels of homophily in order to affect adoption trends. Media influences during later stages, however, were found to be very strong, accounting for a two to four fold increase in the number of people who adopted. This finding is consistent with earlier work that suggests advertising campaigns are enough to confound any word-of-mouth spreading[47].

3.5 Conclusion

This chapter presents descriptive statistics of the spatiotemporal adoption of a web application and proposes a model of technology adoption capable of replicating them. The model extends previous work in two important ways. First, it demonstrates that spatial social networks are crucial to reproducing the dynamics of adoption at a city scale. Second, the model reflects empirical observations that the news volume reacts to the number of adopters with a super-linear trend after a product has reached a

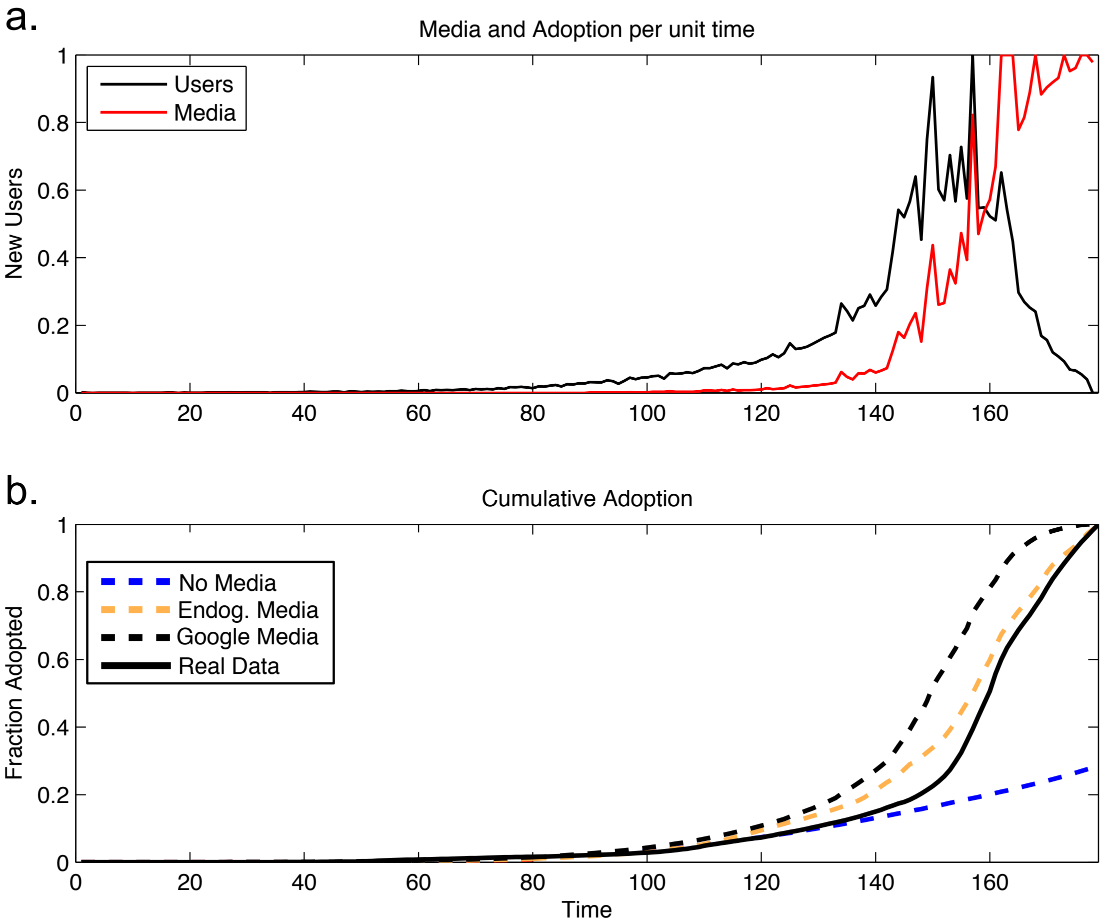

Figure 3-8: **Simulated adoption treating the media as endogenous and increasing with the number of adopters.** (a.) Shows simulated new users per week (normalized to the maximum over the period) as well as normalized media volume each week. (b.) A comparison of all model scenarios is shown. Traditional models, models which do not include media influence are capable of predicting adoption in early periods, but dramatically underestimate total adoption. Including endogenous media effects allows us to make adoption predictions that more closely resemble real data.

critical mass and with random shocks emanating from super-influential people like celebrities or major media events like massive demonstrations.

These results suggest that the model is capable of replicating both micro (at the city level) and macro (at the national level) adoption phenomena and may provide substantial improvement over existing frameworks such as the SI or Bass models. However, some caution is urged in the interpretation of these results. The model may be sensitive to errors in this measurement because simulation relies upon the fraction of a city denoted as early adopters and this fraction was measured empirically from data. While empirical results are intuitive, they may not hold for other products that different from Twitter, such as expensive, durable goods. The model is best applied to goods and services that are very low cost, very easy to tell someone about, and display large positive externalities from a large user base.

In the future, it would be interesting to compare and contrast the spatial diffusion of web apps such as Twitter, with more tangible products such as gadgets, medicine, or cars. For example, it may be possible to use the composition of the cities, as measured from Twitter adoption, to forecast or engineer the adoption of other related kinds of technological innovations. To facilitate further research in this area, a readme and data file has been provided as on-line supplementary material. Empirical data containing city level compositions as well as time series data is available on the web at `http://humnet.scripts.mit.edu/wordpress/2011/06/13/project-modeling-the-diffusion-of-social-contagion/`. Appendix A describes the data in full. This work also represents advances in models of spreading in networks where the roll of demographics, i.e. node attributes, as well as geography is critical for future predictions. These insights may be particularly useful in modeling opinion spreading such as in elections and collective action.

Chapter 4

Big Data and Complex Socio-Technical Systems

This thesis leverages *Big Data* to observe, model, and analyze innovation diffusion involving novel, low cost, and social technologies. The conclusions show that geography and media matter. Looking forward, it is the former result that may receive the most attention. Mobile phones are increasingly equipped with sensors capable of recording their locations. These sensors enable a new spatial dimension for products and services. Activities, tweets, and even coupons are now explicitly tagged in space. Everything becomes local. Moreover, there are roughly six billion mobile phones currently in use. The ubiquity of these devices makes it possible to know the location of nearly every human on the planet at any given time. This fact provokes both concerns and excitement; more significantly it has the potential to radically change our understanding of human behavior.

The same analogies, approaches, and tools utilized thus far in this thesis can also be applied to data from mobile phones and the systems they sense. Each day, billions of people organize themselves in space, interacting with each other and their surroundings. From these movements and actions emerges something that is greater than the sum of its parts – a city. Cities are a personification of the collective actions

[0] Mobile phone statistics provided by the International Telecommunications Union, the United Nations agency for information and communication technologies, `http://www.itu.int/ITU-D/ict/statistics/`.

of residents. They are described not as a sum of these individuals, but rather as separate organisms with unique personalities and characteristics. One city 'never sleeps', while another bathes in 'love and light'. Armed with new sources of data, the next step is to explore the way patterns in the movements and interactions of millions of urban inhabitants contribute to the emergence of urban structure and socioeconomic outcomes.

Just as network models of innovation diffusion were inspired by analogies to physical systems, cities can be viewed a similar lens. Where statistical mechanics seeks to understand the behavior of huge numbers of atoms in a box, urban planning aims to explain how large numbers of people move through a city. In the latter case, the recent explosion of data has created opportunity for research. A better understanding of the way cities function and evolve has the potential to affect billions of people. Earlier this decade, the planet passed an incredible milestone: over half the worlds population became city dwellers[1]. It has been a long march from nomadic tribes and hunter-gathers to the bustling streets of New York, London, and Tokyo, but the gravitational pull of cities has withstood challenges from disease to suburbia. Cities have evolved to house, feed, and entertain billions of inhabitants. They are sustainable systems, using less energy, less water, and producing less waste per-capita than their sprawling alternatives. Cities are centers for business, learning, and culture, facilitating the movement of people and goods from home to work to shops and back again. The complexity and richness of cities has fascinated scholars from fields as diverse as physics and sociology, often inviting more questions than answers when it comes to understanding how they function and how people interact with them.

Mobile phones, now with the ability to pinpoint a user's location to within meters using GPS or WIFI sensors, have the potential to radically improve our knowledge of human mobility patterns within a city. Basic questions about the micro-structure of a city, such as where individuals live and work, that previously could only be answered with small and expensive surveys, can now be explored about a population

[1] Estimates provided by the United Nations Population Fund: http://www.unfpa.org/pds/urbanization.htm.

of millions. Traditionally, features like land use regulations have been determined by an idiosyncratic process of political regulation that provided only snap shots of the ways individuals could use areas of a city. Now, however, it is possible to measure dynamic population density on nearly every street corner at all hours of the day. With these data, the ways people interact with their communities dynamically can be explored and used to inform better solutions to urban planning and transportation issues.

However, in order to realize the potential of this data when applied to complex socio-technical systems, important foundational work must be done. There are a number of important validations and calibrations to consider. Due to technology limitations and privacy concerns, much of these digital data are removed from the context of people and their environment. Whereas traditional surveys and census take great care to collect demographic information from representative samples of the population, data from mobile phones are collected passively from potentially biased sections of the population. Not only must a person own a mobile phone to appear in our data set, but they must also use it. Our window into human mobility is tinted by factors that determine when and where people use digital devices. With this in mind, the first step must be to validate and calibrate data from mobile phones against traditional approaches like travel surveys and census data.

Future research studying urban systems must also reconcile multiple types of mobile phone data with traditional data sources such as zoning regulations, census demographics, and travel surveys. A standard environment should be created in which dynamic mobility data can be layered on top of static indicators to test hypotheses, e.g. whether dynamic human activity in an area is linked to official land use designations. Can activities of mobile phone users be inferred by comparing them to patterns found in travel survey participants? Can human behavior be reliably abstracted from digital breadcrumbs and contextualized with socioeconomic data from different sources? In the broadest sense, the most basic aspects of a city are being measured – where people are, what they are doing, and who they are doing it with. A better understanding of these basic principles can help provide insight into how cities emerge, evolve, and grow.

After the establishing the validity of this data, the possibilities for research are immense. The spatial and temporal resolution coupled with the massive size of these data sets presents a rich opportunity. More accurate measures of the spatial distributions of firms and people may help settle debates about the form and function of cities. Micro-level analysis of mobility patterns, providing real time measurements of population density could be used to infer how land is used dynamically, rather than reliance on static and often antiquated zoning and regulatory data. More accurate models of demand for travel may help transportation planners better position routes and services within a city. Furthermore, knowing how inhabitants move through a city will provide much needed insights into social contact networks used by epidemiologists to model disease spread in urban environments. While these potential contributions are substantial, it is also imperative that this work be performed in a way that ensures the privacy of individuals who generate this data and the companies that capture and store it.

Cutting edge work has used WIFI activity to parse the daily activity patterns of hundreds of college students through mobile phones as well as thousands of campus locations. Behaviors have been decomposed into just a few fundamental patterns that can then be used to differentiate between groups of individuals or types of spaces [17][8]. Similar methods have been applied to data sets on a larger scale, featuring millions of mobile phone users. These reveal that, for all our individual autonomy, humans exhibit highly predictable mobility patterns [44]. The patterns discovered are the first step towards understanding the ways people move across space and interact with each other en masse[48].

Future work will expand upon these results in three important ways. First studies will be scaled from the college campus to entire cities. Second, data will provide insight on how individual's are using these locations dynamically in time rather than merely where people are traveling. Finally, comparing results from multiple data sources in multiple cities will allow researchers to examine biases inherent when studying mobile phone users; large data sets will be placed in the context of real

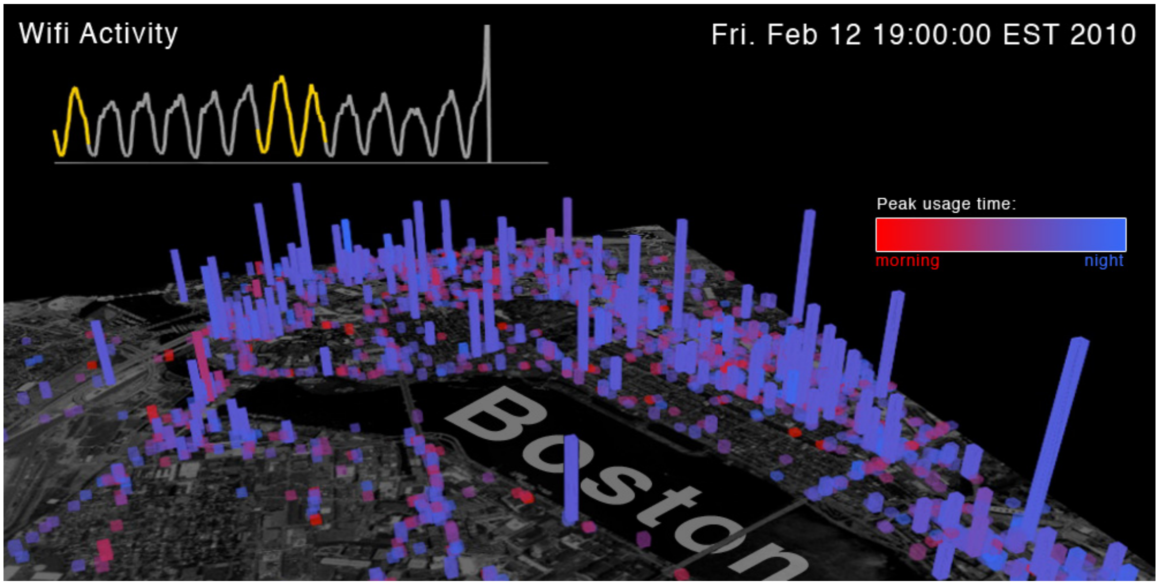

Figure 4-1: **Location based service activity on mobile phones.** This figure displays location based service requests made via smart phones within the city of Boston during an hour of the afternoon. The bars represent the amount of phone activity that occurred on street corners. The color of each bar indicates when that location has the most activity. This type of data is available in hour windows over many months, for nearly every street corner in the city, and for many cities in the world.

demographics.

As an example, consider the visualization in Figure 4-1. The figure shows location based services activity over mobile phones for numerous street corners in Boston on a given hour of the day. The height of each bar represents the amount of phone activity, while the color signifies whether that location has the highest activity in the morning or at night. Next, consider, that this type of data is available for every hour of the day, on every street corner, in every major city. Understanding the patterns in these activity data not only has commercial applications for businesses looking to better understand shopping patterns, but also for urban planners attempting to measure how people move within a city over time.

To conclude, the availability of rich data combined with a willingness to support interdisciplinary approaches promises to propel complex socio-technical systems research to the forefront. It could not be a better time. With a globalized planet facing worldwide problems, transformative solutions on massive scales are necessary to make our lives better and more sustainable. This thesis has provides novel combinations of tools and techniques which leverage big data to understand human behavior. It addresses the future of this domain in the face of ubiquitous technologies that are digital, mobile, and online. This is a time of excitement and imagination - the fun begins.

Appendix A

Dataset

A dataset has been provided to the community a dataset containing empirical data concerning Twitter's adoption. Files are available on the web at `http://humnet.scripts.mit.edu/wordpress/2011/06/13/project-modeling-the-diffusion-of-social-contagi` This data was used to calibrate and test the model of social contagion. It also includes information pertaining to the 408 US cities modeled in a Microsoft Excel workbook with multiple sheets. The data was printed such that the ordering of each sheet (with noted exceptions) is consistent. The first entry in each sheet corresponds to the first city, the second entry to the second city, and so on. The sheet labeled *time_series_week* contains a 180 by 408 matrix where the rows correspond to weeks (time) and the columns to cities. A list of all sheets and their descriptions follows:

1. *city_lat_lon* - contains latitude and longitudinal coordinates for each city.

2. *city_names* - the names and state for each city.

3. *city_type_composition* - the measured fraction of a cities population who were labeled early adopters.

4. *crit_mass_ach_times* - the week at which each city achieved a critical mass (13.5%) of users.

5. *google_news* - weekly news volume has measured from Google Trends (keyword twitter). There are 180 data points, one for each week. Values are normalized

so that the maximum over the interval is 100.

6. *google_search* - weekly search volume as provided by Google Trends (keyword twitter). There are 180 data points, one for each week. Values are normalized so that the maximum over the interval is 100.

7. *time_series_week*- A 180 x 408 matrix where the (i,j)th element corresponds to the number of new uses who signed up for twitter in week i at location j.

8. *total_users_per_city*- the total number of users that signed up for twitter from March 2006 through August 2009 in each city.

Appendix B

City Composition

Table B.1: **Sample cities within each classification (early adopting, late majority, etc.).** Early adopting cities tend to be college towns or have large populations of young, tech-savy users such as Mountain View, CA, while larger metropolitan areas adopted closer to the mean, followed by more rural and remote locations.

Early Adopter	Early Majority	Late Majoirty	Laggards
Total: 60	125	157	66
Ames,IA	Akron,OH	Abilene,TX	Amarillo,TX
Ann-Arbor,MI	Albany,NY	Albright,WV	Beaumont,TX
Arlington,VA	Alexandria,VA	Albuquerque,NM	Bronx,NY
Austin,TX	Alpharetta,GA	Allentown,PA	Cheshire,CT
Beaverton,OR	Amsouth-Bank,TN	Anaheim,CA	Chesterland,OH
Bellevue,WA	Anchorage,AK	Arlington,TX	Clarksville,TN
Bellingham,WA	Anderson,SC	Augusta,GA	Cleveland,GA
Berkeley,CA	Annapolis,MD	Aurora,CO	College-Park,GA
Bethesda,MD	Appleton,WI	Bailey,CO	Columbia,NC
Blacksburg,VA	Asheville,NC	Bakersfield,CA	Columbus,GA

Table B.1 – Continued

Early Adopter	Early Majority	Late Majoirty	Laggards
Bloomington,IN	Athens,OH	Baltimore,MD	Corpus-Christi,TX
Bluefield,VA	Athens-Clarke-County,GA	Baton-Rouge,LA	Detroit,MI
Boston,MA	Atlanta,GA	Bayville,NJ	El-Paso,TX
Boulder,CO	Auburn,AL	Bethlehem,PA	Elk-City,OK
Bozeman,Mt	Bend,OR	Beverly-Hills,CA	England,AR
Cambridge,MA	Boca-Raton,FL	Billings,Mt	Fayetteville,NC
Cary,NC	Boise,ID	Biloxi,MS	Flint,MI
Chapel-Hill,NC	Brooklyn,NY	Birmingham,AL	Fort-Myers,FL
Charlottesville,VA	Burbank,CA	Bowling-Green,KY	Garland,TX
Corvallis,OR	Carlsbad,CA	Bradenton,FL	Grand-Prairie,TX
Davis,CA	Cedar-Rapids,IA	Buffalo,NY	Hamilton,OH
Des-Moines,IA	Champaign,IL	Canton,OH	Hattiesburg,MS
Eugene,OR	Chandler,AZ	Cape-Coral,FL	Hebron,KY
Evanston,IL	Charleston,SC	Charlotte,NC	Jackson,MS
Fairfax,VA	Charleston,WV	Chesapeake,VA	Jacksonville,NC
Franklin,TN	Chattanooga,TN	Cincinnati,OH	Jeffersonton,VA
Grand-Rapids,MI	Chicago,IL	Clearwater,FL	Jupiter,FL
Hoboken,NJ	Chico,CA	College-Station,TX	Kent,WA
Ithaca-College,NY	Cleveland,OH	Colorado-Springs,CO	Killeen,TX
Livermore,CA	Columbia,MO	Columbia,SC	Kissimmee,FL
Madison,WI	Columbus,OH	Dallas,TX	Lake-Charles,LA

Table B.1 – Continued

Early Adopter	Early Majority	Late Majoirty	Laggards
Midwest,WY	Computer Com of Amer,DE	Dayton,OH	Laredo,TX
Minneapolis,MN	Conway,AR	Decatur,GA	Lexington,OK
Missoula,Mt	Coral-Springs,FL	Duluth,MN	Long-Beach,CA
Mountain-View,CA	Costa-Mesa,CA	Durango,CO	Lubbock,TX
Oakland,CA	Denton,TX	Elk-Grove,CA	McAllen,TX
Palo-Alto,CA	Denver,CO	Evansville,IN	Miami,FL
Pasadena,CA	Durham,NC	Everett,WA	Mobile,AL
Portland,ME	East-Lansing,MI	Fayetteville,AR	Modesto,CA
Portland,OR	Easton,PA	Fort-Lauderdale,FL	Montgomery,AL
Provo,UT	Fort-Collins,CO	Fort-Wayne,IN	New-Ringgold,PA
Reston,VA	Frederick,MD	Fort-Worth,TX	Newark,NJ
Rochester,MN	Fredericksburg,VA	Fresno,CA	Newfoundland,PA
Round-Rock,TX	Fremont,CA	Gilbert,AZ	Newport-News,VA
Salt-Lake-City,UT	Frisco,TX	Glendale,AZ	Nokesville,VA
San-Francisco,CA	Fullerton,CA	Glendale,CA	Ocala,FL
San-Jose,CA	Gainesville,FL	Greeley,CO	Palm-Beach,FL
San-Mateo,CA	Greenville,SC	Green-Bay,WI	Palmdale,CA
Santa-Barbara,CA	Gresham,OR	Greensboro,NC	Philippi,WV
Santa-Clara,CA	Harrisburg,PA	Greentown,PA	Portola,CA
Santa-Cruz,CA	Henderson,NV	Greenville,NC	Prosper,TX

Table B.1 – Continued

Early Adopter	Early Majority	Late Majoirty	Laggards
Santa-Monica,CA	Honolulu,HI	Hartford,CT	Queens,NY
Seattle,WA	Huntsville,AL	Hayward,CA	Reading,PA
Silver-Spring,MD	Iowa-City,IA	Heart-Butte,Mt	Shreveport,LA
Somerville,MA	Irvine,CA	Hollywood,FL	Stilwell,OK
St-Paul,MN	Johnson-City,TN	Holtsville,NY	Stockton,CA
State-College,PA	Kalamazoo,MI	Hope,NY	Upper-Marlboro,MD
Sunnyvale,CA	Kansas-City,MO	Houston,TX	Valdosta,GA
Venice,CA	Knoxville,TN	Huntington,WV	Vallejo,CA
West-Lafayette,IN	Lansing,MI	Huntington-Beach,CA	Visalia,CA
	Lawrence,KS	Indianapolis,IN	West-Cornwall,CT
	Lawrenceville,GA	Irving,TX	Whittier,CA
	Leavenworth Lake Wenatchee,WA	Jacksonville,FL	Wilmington,DE
	Lincoln,NE	Jersey-City,NJ	Winston-Salem,NC
	Littleton,CO	Jersey-Shore,PA	Yonkers,NY
	Los-Angeles,CA	Joliet,IL	
	Lynchburg,VA	Kansas-Bank-Amer,KS	
	Manchester,NH	Kansas-City,KS	
	Manhattan,KS	Katy,TX	
	Marietta,GA	Kennesaw,GA	
	Miami-Beach,FL	Kula,HI	

Table B.1 – Continued

Early Adopter	Early Majority	Late Majoirty	Laggards
	Milwaukee,WI	Lafayette,IN	
	Muncie,IN	Lafayette,LA	
	Murfreesboro,TN	Laguna-Beach,CA	
	Napa,CA	Lakeland,FL	
	Naperville,IL	Lancaster,PA	
	New-Haven,CT	Las-Cruces,NM	
	Newark,IL	Las-Vegas,NV	
	North-Hollywood,CA	Lexington,KY	
	Olathe,KS	Little-Rock,AR	
	Olympia,WA	Louisville,KY	
	Omaha,NE	Loveland,OH	
	Orange,CA	Macon,GA	
	Orangeville,UT	Malibu,CA	
	Orlando,FL	Marion,IN	
	Overland-Park,KS	McKinney,TX	
	Petaluma,CA	Melbourne,FL	
	Philadelphia,PA	Melbourne,IA	
	Phoenix,AZ	Memphis,TN	
	Pittsburgh,PA	Mesa,AZ	
	Plano,TX	Midland,TX	
	Pollok,TX	Millersville,MD	
	Raleigh,NC	Monongahela,PA	

Table B.1 – Continued

Early Adopter	Early Majority	Late Majoirty	Laggards
	Redondo-Beach,CA	Morgantown,WV	
	Reno,NV	Murrieta,CA	
	Richmond,VA	Myrtle-Beach,SC	
	Rochester,NY	Naples,FL	
	Salem,OR	New-Brunswick,NJ	
	San-Buenaventura-(Ventura),CA	New-Orleans,LA	
	San-Diego,CA	New-York,NY	
	San-Luis-Obispo,CA	Newark,DE	
	San-Marcos,TX	Newport-Beach,CA	
	Santa-Clarita,CA	Norman,OK	
	Santa-Fe,NM	Oceanside,CA	
	Santa-Rosa,CA	Oklahoma-City,OK	
	Sarasota,FL	Orange,TX	
	Scottsdale,AZ	Palm-Springs,CA	
	Sioux-Falls,SD	Panama-City,FL	
	South-Bend,IN	Pensacola,FL	
	Spokane,WA	Peoria,AZ	
	Springfield,IL	Peoria,IL	
	St-Louis,MO	Piatt,PA	

Table B.1 – Continued

Early Adopter	Early Majority	Late Majoirty	Laggards
	Stamford,CT	Pinckney,MI	
	Stillwater,OK	Providence,RI	
	Tempe,AZ	Puyallup,WA	
	Thousand-Oaks,CA	Rancho-Cucamonga,CA	
	Tulsa,OK	Redding,CA	
	Tustin,CA	Riverside,CA	
	Vancouver,WA	Roanoke,VA	
	Washington,DC	Rockford,IL	
	West-Hollywood,CA	Roseville,CA	
	Williamsburg,VA	Sacramento,CA	
	Wilmington,NC	San-Antonio,TX	
	Winter-Park,FL	San-Bernardino,CA	
	Woonsocket,RI	San-Clemente,CA	
		Savannah,GA	
		Scranton,PA	
		Siloam-Springs,AR	
		Simi-Valley,CA	
		Spartanburg,SC	
		Springfield,MO	
		St-Augustine,FL	
		St-Petersburg,FL	
		St-Stephen,SC	

Table B.1 – Continued

Early Adopter	Early Majority	Late Majoirty	Laggards
		Staten-Island,NY	
		Sugar-Land,TX	
		Surprise,AZ	
		Syracuse,NY	
		Tacoma,WA	
		Tallahassee,FL	
		Tampa,FL	
		Temecula,CA	
		Toledo,OH	
		Topeka,KS	
		Torrance,CA	
		Traverse-City,MI	
		Tucson,AZ	
		Tuscaloosa,AL	
		Tyler,TX	
		Virginia-Beach,VA	
		Waco,TX	
		West-Chester,PA	
		West-Palm-Beach,FL	
		Winchester,VA	
		Woodbridge,VA	
		Worcester,MA	
		Young,AZ	

Appendix C

Model Implementation

The model was implemented in Python utilizing the open source SciPy and NumPy libraries to perform calculations and statistics. Because the simulation is stochastic, multiple runs were performed for each parameter set. To efficiently generate ensembles and sweep the parameter space, a special Model class was implemented. The class contains two data fields, each containing an instance of another custom class. The first data field is reserved for a Params class, storing the model parameters for that run. The second is a Support class, storing data from the simulation. Each new set of model parameters is a new Model Object. Each run for the same set of parameters is run under the same Model Object, but results are exported as its own text file. Upon initialization of a Model Object, parameters are set to default values and memory is allocated for storing input and output data. The Model class has a number of functions that implement the procedure described in Section 3.3. For example, the *InitNetwork* function within the Model class initializes a population of agents and connects them in a social network based on parameters input to the Model Object. These class functions also include exporting features that output simulation results to text files for later analysis. Moreover, because each instance of a Model Object is a self-contained simulation, multiple runs can be parallelized, significant reducing computation time. The remainder of this section describes the input parameters of the Model and lists class functions.

C.1 Model Class

The following sections list the functions available to the Model Object.

C.1.1 *class* **model**

model.**params**

A data field that stores a Params Object containing the parameter settings for the model's configuration. The Params class is documented in detail below.

model.**support**

A data field that stores simulation data and output. A more detailed description of the data stored can be found below.

model.**init_params(model obj)**

Takes a model object as an input and sets params and support data fields to default values.

model.**init_network(model obj)**

Initializes a population of agents and constructs a social network based on parameters of the model.

model.**simulate(model obj)**

Simulates the diffusion of an innovation based on the parameters of the Model Object.

model.**export_run(model obj)**

Exports the results from numerical simulation to text files.

model.**run(model obj, integer)**

Begins the initiation and simulation process. The second parameters indicates the run number for the given parameters setting.

C.2 Model Parameters

Each instance of a Model Object contains a data field for a Params Object. This Params Object stores the model settings for a given run. The following section lists the fields present in the Params class.

C.2.1 *class* params

params.**T** *type:* Integer

The number of time steps to run the simulation for. In most cases, this value represented weeks.

params.**L** *type:* Integer

The number of cities in which agents can be placed. In the case of Twitter, 408 cities were simulated.

params.**N** *type:* Integer

Total number of nodes in the network.

params.**AVGS** *type:* Integer

Number of runs to average over for each ensemble at a particular parameters setting.

params.**RUNS** *type:* Integer

The number of runs to be performed for each parameter setting

params.**Kavg** *type:* Double $[0, \infty)$

Average degree of the social network.

params.**Pref** *type:* Double $[0, 1]$

The probability of a given friendship being between two nodes of the same type.

params.**Allowed** *type:* Double $[0, 1]$

The percent of all links allowed to exist between nodes of a different type.

params.**Sus** *type:* (Double, Double) $[0, 1]$

Stored as a data pair, the first value refers to the susceptibility of regular adopters, while the second corresponds to early adopters.

params.**Seed** *type:* Integer

The number of nodes initially using the innovation.

params.**MEDIA** *type:* Boolean

A boolean value indicating if the media is present or not.

params.**a** *type:* Double $[0, 1]$

The probability that an agent heads the media's recommendation and adopts during a period.

params.**g** *type:* Double $[0, \infty)$

The exponential power of the media's growth rate.

params.**TIME** *type:* String

The time a particular model run was initiated (for data storage purposes).

params.**PLOT** *type:* Boolean

A boolean indicating whether results will be plotted or not.

params.**EXPORT** *type:* Boolean

A boolean indicating whether results will be exported or not.

params.**BASEPATH** *type:* String

The base file path.

params.**PATH** *type:* String

A more specific file path to output folders.

params.**NET** *type:* "geographic" or "random"

The type of social network created, spatially embedded or not.

params.**POP** *type:* [Integer, . . . , Integer]

The population of each city. Can be inputed from a file containing empirical populations of Twitter users or set arbitrarily.

params.**QUIET** *type:* Boolean

A boolean suppressing consol output.

params.**FIT** *type:* Boolean

A boolean indicating whether the model will asses the fit of its simulation to real data.

C.3 Model Data

The data from each simulation is stored in a custom data class. This class stores all information required to numerically simulation adoption and can be exported to text files after each run for later analysis.

C.3.1 *class* **support**

support.**Nodes** *type:* Array

Contains an array storing all agents in the network. Each agent is stored as a dictionary containing the following attributes: id, location, infected status, type, neighbors.

support.**Status** *type:* Array

An array containing the infection status of each agent. Susceptible agents have status 0 while infected agents are status 1. Summing all elements in this array gives the total number of infected individuals

support.**Locs** *type:* Array

An array storing the city location of each agent.

support.**Degs** *type:* Array

An array storing the degree of each agent in the network.

support.**Poplist** *type:* [[Array],. . . ,[Array]]

A list of arrays. Each array in the list contains the ids of all agents placed in that city.

support.**Pops** *type:* Array

An array containing the total population of each city in the simulation.

support.**Comp** *type:* Array

An array containing the fraction of each city's population of the early adopting type. The remaining fraction is the percentage of each cities population of the regular type.

support.**Type** *type:* Array

An array containing the type of each agent in the population.

support.**Coords** *type:* [(double, double), . . . , (double, double)]

Coordinates of cities. Used for plotting purposes only.

support.**Names** *type:* [String, . . . , String]

Array of strings containing city names. Used for plotting purposes only.

support.**D** *type:* Array[][]

An $L \times L$ array, where L is the number of locations, of the euclidian distance between cities. For example, support.D[i][j] returns the distance between cities i and j.

support.**PDF** *type:* Array

An array containing a numerical approximation of the probability density function of choosing a friend in a city a distance r away. The accuracy of this approximation can be changed by decreasing the interval between elements. Two pdfs were used for this thesis. The first was uniform over the maximum distance between two cities. The second was the power law $p_r = r^{-1.2} + \nu$, motivated by empirical results from Liben-Nowell et. al's[33] study of an online social network.

support.**CDF** *type:* String

An array containing a numerical approximation of the cumulative density function of choosing a friend in a city a distance r away. The accuracy of this approximation can be changed by decreasing the interval between elements. Two cdfs were used for this thesis. The first was uniform over the maximum distance between two cities. The second was the power law $p_r = r^{-1.2} + \nu$, motivated by empirical results from Liben-Nowell et. al's[33] study of an online social network.

support.**dx** *type:* Array

The spatial resolution of the distance function used as input to the support.PDF and support.CDF variables.

support.**TS** *type:* [[Array],. . . ,[Array]]

A list of arrays containing the time series of adoption for each individual city.

support.**AggTS** *type:* Array

An array containing the time series of aggregate, national level adoption.

support.**EarlyAdopt** *type:* Array

A array storing the time series of the number of agents of type early adopter that adopted each period.

support.**RegAdopt** *type:* Array

A array storing the time series of the number of agents of type early adopter that adopted each period.

support.**Pks** *type:* Array

An array containing the period in which each city reached critical mass.

support.**PkError** *type:* Array

An array containing the different between the simulated time of critical mass for each city and the time measured in real data (if real data is available).

support.**RunError** *type:* Array

An array containing the average error in critical mass achievement time across all cities for each run at a constant set of model parameters.

support.**Run** *type:* Integer

The index of the particular run of the model for a constant set of parameters. Due to the stochastic nature of the algorithm, each parameter settings are run multiple times then statistics are performed on the ensemble average.

support.**M** *type:* Array

An array containing a time series of mass media influence.

support.**real_pops** *type:* Array

An array containing measurements of real city populations if data is available as input.

support.**real_comps** *type:* Array

An array containing measurements of real city compositions (e.g. fraction of early adopters) if data is available as input.

support.**real_peaks** *type:* Boolean

An array containing measurements of real city critical mass achievement times if data is available as input.

C.4 Agent Class

A custom class is also created for agents. Each Agent object has a number of associated data fields.

C.4.1 *class* **agent**

agent.**ID** *type:* Integer

An unique integer used as an ID for the agent.

agent.**loc** *type:* Integer

The ID of the city in which that agent is located.

agent.**type** *type:* Integer

The type of the agent, 0 corresponding to early adopter, 1 to regular adopter.

agent.**sus** *type:* Double

The susceptibility of the agent. This is determined by type, early or regular adopter. The value of this parameter is interpreted as the probability an agent will adopt an innovation is asked by a neighbor.

agent.**deg** *type:* Integer

The degree of an agent. This value is pulled from a Poisson distribution whose average can be set in the parameters of the model.

agent.**nbrs** *type:* Array

An array of integers corresponding to the IDs of all other nodes in the network connected to that particular agent.

agent.**status** *type:* Integer

> The status of an agent. A value of 0 denotes that the agent is susceptible to an innovation, while 1 refers to agents who have already adopted.

C.5 Run Controller

To efficiently simulate the diffusion of innovation and analyze the results, a controller was written to sweep various parameter ranges and perform a number of runs at each unique parameter settings. This controller also introduced parallelization so that multiple runs and parameter settings could be simulated at once, greatly reducing computation times.

The controller creates a new process for each unique set of parameters. The total number of concurrent processes is limited by the number of CPUs present in the machine. In general, a single parameter set is simulated by a single process. The controller begins by creating a new instance of a Model object and initializing it to the parameter set of that model. The agent population and social network is then created using the methods of the Model class. After the population has been created, adoption is simulated under the parameter values currently being tested. A single Model object runs multiple simulations for its unique parameter setting. The output from each simulation as well as a list of the parameters the simulation was run at are then saved as text files in a folder labeled by the parameter settings. This architecture allows for parameter sweeps to be parallelized, reducing computation times. If it is only a single parameter setting being tested, each individual run can further be spread out onto different processes and CPUs.

C.6 Simulation Algorithms

Algorithms were written to efficiently create social networks and simulation diffusion of innovations on them. This section describes, in pseudocode, the implementation of the simulation initialization and dynamics.

Algorithm C.6.1 describes the creation of a random network that does not consider geography. The first step initializes an agent population, allocating memory for N nodes. The degree of each node is also chosen according to some distribution. In most cases, a Poisson degree distribution was used. In their initialized states, agents can be thought of as nodes in a network with a set number of stubs. Stubs from two different agents are connected to form a link in the social network. For each node in the population, the algorithm first checks to make sure there is at least one available stub. If there is an opening, a neighbor is chosen. To control homophily in the network, the agent chooses a friend of the same type with a certain probability. If there is no homophily, the neighbor is chosen at random. A check is also performed to make sure that the chosen neighbor also has unused stubs. After a suitable neighbor is found, the ID of the new friend is added to the neighbor list of the current node and the current node's ID is added to the list of the new friend. This process is repeated until the current node has filled all available stubs. The algorithm then moves to the next agent in the population and performs the same matching procedure. As the number of available stubs become small, it may be impossible to match a node with a suitable neighbor. In these cases, any random available stub is chosen. In practice, this happens only a small number of times and is insignificant with sufficiently large populations.

Algorithm C.6.1: RANDOM_NETWORK($model.params, model.support$)

```
$agent\_pop \leftarrow init\_nodes();$
for $i \leftarrow 0$ to $params.N$
  do while $length(node_i.nbrs) < node_i.deg$
    if $random() > params.PREF$
      then $new\_nbr \leftarrow$ random agent that can accept link and is different type

      else $new\_nbr \leftarrow$ random agent that can accept link and is same type
    $node_i.nbrs.append(new\_nbr)$
    $new\_nbr.nbrs.append(node_i)$
```

Algorithm C.6.2 describes a slightly more complicated procedure for generating geographically biased social networks. An agent population is initialized as before and each agent's connections are assigned. If an agent has an open stub, a random geographic distance, d, is chosen from a probability function specified in the parameters of the model. For example, to replicate the empirically measured probability, p_r, that two individuals, separated by a distance r are friends, this was pdf was set to a power law. In practice, any distribution can be used. However, agents are not dispersed continuously through space. They are placed in cities which have set locations. A city, l, is chosen such that the distance between the location of the current node and l is minimized. Next, the current node then chooses a suitable new neighbor from that city, l. This new neighbor must have available stubs and must be the correct type if homophily is present. If a match is made, the neighbor lists of the current node and of the new neighbor are updated. This process is repeated until all of the current nodes empty stubs are filled. The algorithm then performs the same procedure for the next agent, continuing until all connections are made. In this case, cities run out of suitable nodes faster than the entire network does. If an agent is unable to find a suitable neighbor in a chosen city, l, the second city closest to a distance, d, away is chosen. If too many attempts are made, a random neighbor is assigned. The networks generated by this algorithm were tests to ensure they reproduced to the empirical distributions measured in real networks.

Algorithm C.6.2: GEOGRAPHIC_NETWORK(*model.params*, *model.support*)

```
$agent\_pop \leftarrow init\_agents();$
for $i \leftarrow 0$ to $params.N$
  do while $length(agent_i.nbrs) < agent_i.deg$
    $d \leftarrow random\_dist()$
    comment: $random\_dist()$ returns a distance from the probability distribution $params.PDF$.
    $l \leftarrow$ city closest to a distance, $d$, from $agent_i.city$
    if $random() > params.PREF$
      then $new\_nbr \leftarrow$ random available agent from city $l$ of different type
      else $new\_nbr \leftarrow$ random available agent from city $l$ of same type
    $agent_i.nbrs.append(new\_nbr)$
    $new\_nbr.nbrs.append(agent_i)$
```

Algorithm C.6.3 details the dynamic simulation of innovation diffusion. Innovation diffusion is simulated after the agent population is initiated and the social network is grown. At first, no one has adopted the new technology, so the process must be seeded. This is done by changing the status of a small fraction of the population (less than 0.001%) to infected. After diffusion is seeded, time proceeds in discrete steps. In general, any length time period can be used, but for the majority of simulations in this thesis, each period was interpreted as a week. In each period, an array is created with the IDs of all the currently infected, currently susceptible, and currently at-risk agents. At-risk agents are susceptible agents who are connected to an infected neighbor. It is important to note that the elements of the at-risk list are not unique. If a susceptible agent has three infected neighbors, that agent will appear in the at-risk list three times. This is because each infected neighbor recommends the susceptible agent adopt in each period. The more infected neighbors a node has, the higher the probability is that susceptible agent adopts.

The first type of adoption that can occur is due to the word-of-mouth mechanism. An agent hears about an innovation from a friend, then decides if it will adopt. This is simulated by iterating over the at-risk list. Each at-risk agent flips a biased coin every time it appears in the list to determine if it will adopt. This coin is represented as a random number generator. The probability that the coin lands on 1 (adopt) is equal to the susceptibility of agent's type. If this occurs, the agent's status is changed to 1, infected. Otherwise, it remains 0, susceptible. Early adopting types have a higher probability of adopting than regular. If no media is present, the period ends and time series are updated to reflect the number of agents who adopted that period. The time series are further broken down by type of agent and the city an agent adopted from. The procedure then repeats itself, starting by creating updated currently-infected and currently-susceptible lists.

If the media is present, however, more adoption can occur after word-of-mouth diffusion is simulated. First, the strength of the media is calculated. As outlined in Section 3.3, media volume is endogenous, depending on the number of people who have already adopted an innovation. For each period, the fraction of the total agent population who has already adopted is calculated. This fraction is then raise to some power, model parameter *params.g*, reflecting the non-linear relationship observed in empirical data. Finally, a random shock, ϵ is added. In total the media volume in period t is given by $M(t) = I(t)^{\gamma} + \epsilon$. In addition to media volume, there is also the susceptibility of each agent to listen to the media's message. In each period, the currently susceptible list is iterated over and each agent flips a coin, adopting with probability $\alpha \cdot M(t)$ and remaining susceptible otherwise. After all susceptible agents have flipped a coin, the time series are updated as in the case with no media.

Algorithm C.6.3: SIMULATIONS($model.params, model.support$)

```
$seed\_infection()$
for $t \leftarrow 0$ to $params.T$
do
    $currently\_infected \leftarrow$ [IDs of all infected agents at time $t$]
    $currently\_susceptible \leftarrow$ [IDs of all susceptible agents at time $t$]
    $at\_risk \leftarrow$ [susceptible neighbors of all infected nodes $t$]
    comment: The $at\_risk$ array includes duplicates of nodes with multiple infected friends.
    comment: First spread adoption via word-of-mouth.
    for each $agent \in at\_risk$
        $r \leftarrow random()$
        if $r < agent.sus$ : $agent.status \leftarrow 1$
          else : $agent.status \leftarrow 0$
    comment: If media is present, calculate its volume.
    if $params.MEDIA == True$
        $\epsilon \leftarrow random()$
        $support.M(t) \leftarrow [\frac{length(currently\_infected)}{params.N}]^{params.g} + \epsilon$
        for each $agent \in currently\_susceptible$
            $r \leftarrow random()$
            if $r < (params.a) \cdot (support.M(t))$ : $agent.status \leftarrow 1$
              else : $agent.status \leftarrow 0$
    $update\_timeseries()$
```

Bibliography

[1] A. Allaway. Spatial diffusion of a new loyalty program through a retail market. *Journal of Retailing*, 79(3):137–151, 2003.

[2] Brian W. Arthur and David A. Lane. Information Contagion. *Structural Change and Economic Dynamics*, 4(1):81–104, June 1993.

[3] Eytan Bakshy, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 65–74, New York, NY, USA, 2011. ACM.

[4] Duygu Balcan, Bruno Gonçalves, Hao Hu, José J. Ramasco, Vittoria Colizza, and Alessandro Vespignani. Modeling the spatial spread of infectious diseases: The GLobal Epidemic and Mobility computational model. *Journal of Computational Science*, 1(3):132–145, August 2010.

[5] Frank M. Bass. A New Product Growth for Model Consumer Durables. *Management Science*, 15(5):215–227, January 1969.

[6] Matthew A. Baum and Angela S. Jamison. The Oprah Effect: How Soft News Helps Inattentive Citizens Vote Consistently. *The Journal of Politics*, 68(04):946–959, November 2006.

[7] Ronald S. Burt. Social Contagion and Innovation: Cohesion Versus Structural Equivalence. *American Journal of Sociology*, 92(6):1287–1335, 1987.

[8] Francesco Calabrese, Jonathan Reades, and Carlo Ratti. Eigenplaces: Segmenting Space through Digital Signatures. *IEEE Pervasive Computing*, 9(1):78–84, 2010.

[9] Claudio Castellano, Santo Fortunato, and Vittorio Loreto. Statistical physics of social dynamics. *Reviews of Modern Physics*, 81(2):591–646, May 2009.

[10] D. Centola, V. Eguiluz, and M. MacY. Cascade dynamics of complex propagation. *Physica A: Statistical Mechanics and its Applications*, 374(1):449–456, January 2007.

[11] Damon Centola. Failure in Complex Social Networks. *The Journal of Mathematical Sociology*, 33(1):64–68, 2009.

[12] Damon Centola. The Spread of Behavior in an Online Social Network Experiment. *Science*, 329(5996):1194–1197, September 2010.

[13] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and Krishna P. Gummadi. Measuring User Influence in Twitter: The Million Follower Fallacy. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM).*

[14] Klaus Dietz. Epidemics and Rumours: A Survey. *Journal of the Royal Statistical Society. Series A (General)*, 130(4), 1967.

[15] P. S. Dodds and D. J. Watts. A generalized model of social and biological contagion. *Journal of Theoretical Biology*, 232(4):587–604, February 2005.

[16] Robin I. M. Dunbar. The social brain hypothesis. *Evol. Anthropol.*, 6(5):178–190, 1998.

[17] Nathan Eagle and Alex Pentland. Eigenbehaviors: identifying structure in routine. *Behavioral Ecology and Sociobiology*, 63(7):1057–1066, May 2009.

[18] James H. Fowler. The Colbert Bump in Campaign Donations: More Truthful than Truthy. *PS: Political Science & Politics*, 41(03):533–539, 2008.

[19] Tal Garber, Jacob Goldenberg, Barak Libai, and Eitan Muller. From Density to Destiny: Using Spatial Dimension of Sales Data for Early Prediction of New Product Success. *MARKETING SCIENCE*, 23(3):419–428, January 2004.

[20] Bruno Gonalves, Nicola Perra, and Alessandro Vespignani. Modeling users' activity on twitter networks: Validation of dunbar's number. *PLoS ONE*, 6(8):e22656, 08 2011.

[21] Mark Granovetter. Threshold Models of Collective Behavior. *American Journal of Sociology*, 83(6):1420–1443, 1978.

[22] Mark S. Granovetter. The Strength of Weak Ties. *American Journal of Sociology*, 78(6):1360–1380, 1973.

[23] Roger M. Heeler and Thomas P. Hustad. Problems in Predicting New Product Growth for Consumer Durables. *Management Science*, 26(10):1007–1020, October 1980.

[24] Benjamin F. Jones, Stefan Wuchty, and Brian Uzzi. Multi-University Research Teams: Shifting Impact, Geography, and Stratification in Science. *Science*, 322(5905):1259–1262, November 2008.

[25] Brian Karrer and M. E. J. Newman. Message passing approach for general epidemic models. *Physical Review E*, 82(1):016101+, July 2010.

[26] Elihu Katz. The Two-Step Flow of Communication: An Up-To-Date Report on an Hypothesis. *The Public Opinion Quarterly*, 21(1):61–78, 1957.

[27] Michael L. Katz and Carl Shapiro. Network Externalities, Competition, and Compatibility. *The American Economic Review*, 75(3):424–440, 1985.

[28] Jon Kleinberg. Cascading behavior in networks: Algorithmic and economic issues, 2007.

[29] David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-Lászlo Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. Computational Social Science. *Science*, 323(5915):721–723, February 2009.

[30] Jure Leskovec, Lada A. Adamic, and Bernardo A. Huberman. The dynamics of viral marketing. *ACM Trans. Web*, 1, May 2007.

[31] Jure Leskovec, Lars Backstrom, and Jon Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 497–506, New York, NY, USA, 2009. ACM.

[32] David Liben-Nowell and Jon Kleinberg. Tracing information flow on a global scale using Internet chain-letter data. *Proceedings of the National Academy of Sciences*, 105(12):4633–4638, March 2008.

[33] David Liben-Nowell, Jasmine Novak, Ravi Kumar, Prabhakar Raghavan, and Andrew Tomkins. Geographic routing in social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 102(33):11623–11628, August 2005.

[34] Dunia López-Pintado and Duncan J. Watts. Social Influence, Binary Decisions and Collective Dynamics. *Rationality and Society*, 20(4):399–443, November 2008.

[35] M. E. J. Newman, I. Jensen, and R. M. Ziff. Percolation and epidemics in a two-dimensional small world. *Physical Review E*, 65(2):021904+, January 2002.

[36] Jukka-Pekka Onnela, Samuel Arbesman, Albert-László Barabási, and Nicholas A. Christakis. Geographic constraints on social network groups. November 2010.

[37] Jukka-Pekka Onnela and Felix Reed-Tsochas. Spontaneous emergence of social influence in online systems. *Proceedings of the National Academy of Sciences*, 107(43):18375–18380, October 2010.

[38] Andrew Pease and Paul R. Brewer. The Oprah Factor: The Effects of a Celebrity Endorsement in a Presidential Primary Campaign. *The International Journal of Press/Politics*, 13(4):386–400, October 2008.

[39] Hazhir Rahmandad and John Sterman. Heterogeneity and Network Structure in the Dynamics of Diffusion: Comparing Agent-Based and Differential Equation Models. *Management Science*, 54(5):998–1014, May 2008.

[40] Everett M. Rogers and Everett Rogers. *Diffusion of Innovations, 5th Edition*. Free Press, 5th edition, August 2003.

[41] D. M. Romero, B. Meeder, and J. Kleinberg. Differences in the Mechanics of Information Diffusion Across Topics: Idioms, Political Hashtags, and Complex Contagion on Twitter. 2010.

[42] Matthew J. Salganik, Peter S. Dodds, and Duncan J. Watts. Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market. *Science*, 311(5762):854–856, February 2006.

[43] Thomas C. Schelling. Hockey Helmets, Concealed Weapons, and Daylight Saving: A Study of Binary Choices with Externalities. *The Journal of Conflict Resolution*, 17(3):381–428, 1973.

[44] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of Predictability in Human Mobility. *Science*, 327(5968):1018–1021, February 2010.

[45] Jeffrey Travers and Stanley Milgram. An Experimental Study of the Small World Problem. *Sociometry*, 32(4):425–443, 1969.

[46] Thomas W. Valente. *Network Models of the Diffusion of Innovations (Quantitative Methods in Communication Subseries)*. Hampton Press (NJ), January 1995.

[47] Christophe Van den Bulte and Gary L. Lilien. Medical Innovation Revisited: Social Contagion versus Marketing Effort. *American Journal of Sociology*, 106(5):1409–1435, March 2001.

[48] Pu Wang and Marta C. González. Understanding spatial connectivity of individuals with non-uniform population density. *Physical and Engineering Sciences*, 367(1901):3321–3329, August 2009.

[49] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, June 1998.

[50] Duncan J. Watts. A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences*, 99(9):5766–5771, April 2002.

[51] Duncan J. Watts and Peter S. Dodds. Influentials, Networks, and Public Opinion Formation , 2007.

[52] Duncan J. Watts, Roby Muhamad, Daniel C. Medina, and Peter S. Dodds. Multiscale, resurgent epidemics in a hierarchical metapopulation model. *Proceedings of the National Academy of Sciences of the United States of America*, 102(32):11157–11162, August 2005.

[53] F. Wu, B. Huberman, L. Adamic, and J. Tyler. Information flow in social groups. *Physica A: Statistical and Theoretical Physics*, 337(1-2):327–335, June 2004.

[54] Jaewon Yang and Jure Leskovec. Patterns of temporal variation in online media. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 177–186, New York, NY, USA, 2011. ACM.